



Northeastern University

Customer Segmentation and Behavior Predictions

Submitted by:

Margi Shah

Mohan Bhosale

Shruti Suhas Kute

Under the Guidance of

Prof. David Brady

Contents

1. Summary.....	3
1.1 Problem Statement.....	3
1.2 Project Objectives.....	3
1.3 Project Introduction.....	4
2. EDA.....	5
2.1 Dataset Description.....	6
2.2 EDA.....	7
3. Frequent Itemset and Association Rule Mining.....	11
3.1 Apriori Algorithm.....	11
3.2 Transaction Matrix Representation.....	12
3.3 Product Recommendation from Association Rules.....	13
4. Customer Segmentation Using Clustering.....	15
4.1 Clustering Metrics.....	15
4.2 GMM.....	16
4.3 RFM Analysis.....	18
4.4 K-means.....	19
4.5 Customer Segmentation and Visualization.....	21
5. Link Analysis.....	23
5.1 Link Analysis and Community Detection.....	23
5.2 Degree Centrality Analysis.....	24
5.3 Betweenness Centrality Analysis.....	25
5.4 Closeness Centrality Analysis.....	26
5.5 Eigenvector Centrality Analysis.....	27
6. Hybrid Recommender with Dimensionality Reduction.....	28
6.1 Hybrid Recommender System.....	28
6.2 Dimensionality Reduction using SVD.....	29
7. Conclusion.....	30
8. Usage.....	30

1. SUMMARY

1.1 Problem Statement

E-commerce platforms generate immense amounts of transactional data daily, capturing details about customer purchases, browsing habits, and preferences. However, transforming this raw data into actionable insights that can enhance customer experiences and boost revenue remains a complex challenge. Businesses need to go beyond traditional analysis methods to uncover deeper patterns, such as identifying nuanced customer behavior trends and understanding the relationships between products. For instance, insights into frequently purchased product combinations can drive targeted marketing strategies and boost cross-selling opportunities.

1.2 Project Objectives

Key goals of this project include leveraging advanced data analytics techniques to extract actionable insights from customer purchasing data, ultimately driving better business decisions. One primary objective is to identify frequent product combinations and generate association rules that enable businesses to design effective cross-selling and upselling strategies. By understanding which products customers commonly purchase together, businesses can create targeted promotional offers, bundled deals, and personalized suggestions that maximize transaction value while enhancing the overall shopping experience.

Key Goals include:

- To identify frequent product combinations and generate association rules for effective cross-selling and upselling strategies.
- To segment customers into meaningful groups using clustering methods, such as Gaussian Mixture Models, to enable targeted marketing campaigns.
- To apply dimensionality reduction techniques for managing high-dimensional customer-product interaction data efficiently.
- To construct and analyze a product co-purchase network to understand product relationships and identify influential products using social network and link analysis.
- To develop and evaluate a recommender system that suggests personalized product recommendations based on historical purchase patterns.

Overall, this comprehensive analysis is designed to transform raw transactional data into actionable business strategies. By identifying product relationships, segmenting customers effectively, and developing advanced recommendation systems, the project aims to enhance marketing strategies, improve customer satisfaction, and optimize business outcomes. These insights equip businesses with the tools they need to remain competitive in a rapidly evolving marketplace.

1.3 Project Introduction

This project focuses on analyzing customer purchasing patterns to uncover key trends and behaviors that enable businesses to optimize their marketing strategies, improve inventory management, and drive sales growth. By leveraging advanced data analytics techniques, the project identifies actionable insights that help businesses understand their customers more deeply. Through the segmentation of customers into distinct groups, the discovery of product associations, and the analysis of seasonal trends, this approach empowers organizations to enhance personalized marketing, develop effective product bundling strategies, and forecast demand with greater accuracy.

Customer segmentation plays a pivotal role in understanding diverse buying behaviors. By grouping customers based on factors such as purchase frequency, average spending, and product preferences, businesses can tailor their marketing efforts to cater to specific customer needs. Furthermore, uncovering product associations—such as items frequently purchased together—provides opportunities for cross-selling and bundling. For example, identifying that customers who purchase sports apparel often buy fitness accessories can lead to bundled deals that not only increase sales but also enhance customer satisfaction.

To achieve these insights, the project employs advanced techniques like clustering, dimensionality reduction, and association rule mining. Clustering algorithms group customers into meaningful categories, while dimensionality reduction simplifies complex datasets to reveal key patterns. Additionally, association rule mining identifies frequently purchased product combinations, offering businesses a foundation to design targeted promotions and optimize inventory management. This comprehensive analysis enables businesses to align their strategies with customer behavior, ensuring long-term success in an increasingly data-driven marketplace.

2. EDA

2.1 Dataset Description

This project leverages the publicly available **Online Retail II** dataset, a comprehensive collection of transactional data sourced from a UK-based non-store online retailer. The dataset spans a two-year period, capturing transactions between **December 2009** and **December 2011**, providing a detailed account of customer purchasing activities. It primarily focuses on purchases of unique, all-occasion giftware, catering mostly to wholesalers. The dataset includes a wealth of information, such as invoice numbers, stock codes, descriptions, quantities purchased, customer IDs, and geographic details, offering a robust foundation for exploring patterns in consumer behavior.

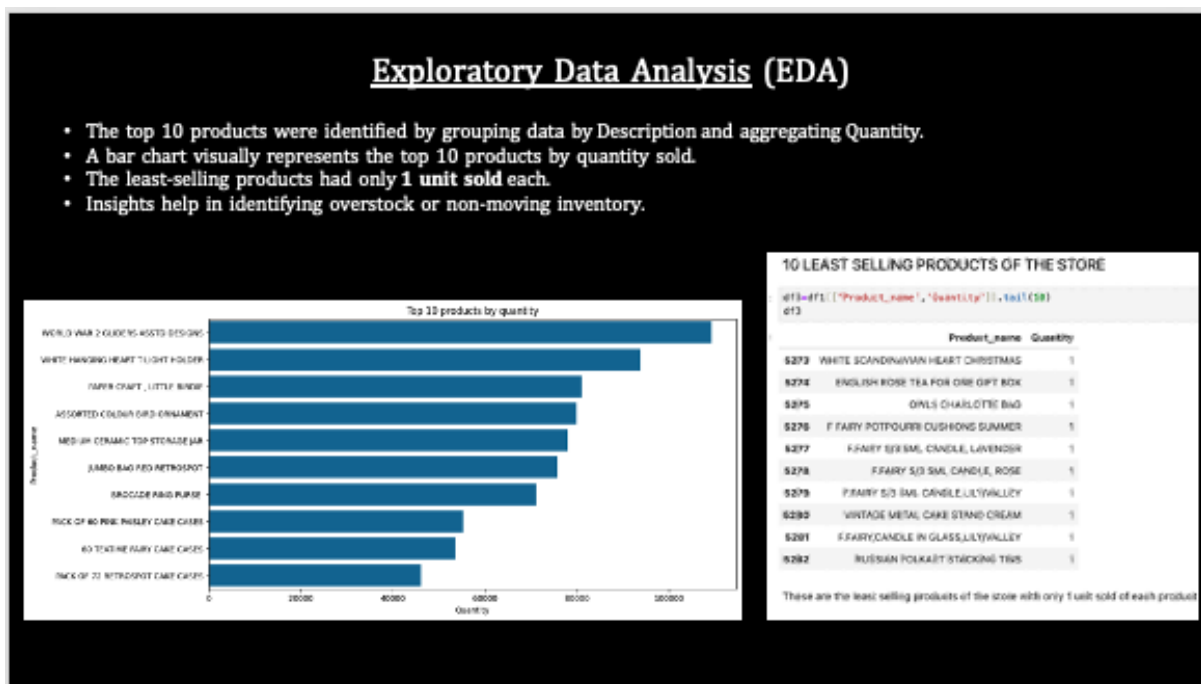
The richness and diversity of the dataset make it an ideal candidate for deriving actionable insights into customer purchasing patterns and designing data-driven marketing strategies. It reflects real-world transactional behaviors, making it possible to uncover trends, seasonal demand, and associations between products that are frequently purchased together. For example, by analyzing co-purchase data, businesses can identify opportunities for product bundling and targeted promotions, driving increased sales and customer satisfaction.

Variable	Description
InvoiceNo	Unique identifier for each transaction, with cancellations marked by an initial 'C'
StockCode	Unique identifier for each product sold.
Description	Text description of products, providing insights in customer preferences.
Quantity	Number of units purchased per transaction.
InvoiceDate	Timestamp of each transaction, enabling temporal analyses.
UnitPrice	Product price per unit in GBP , supporting revenue and sales analysis.
CustomerID	Unique identifier for each customer, facilitating customer segmentation.
Country	Unique identifier for each customer, facilitating customer segmentation.

Table 1. Captures Dataset Attributes

The dataset initially comprised 1,067,371 rows and 8 columns, but it required extensive preprocessing to ensure data integrity and relevance for analysis. Missing values were identified in key columns, including 243,007 rows with missing **Customer ID** values and 4,382 rows missing **Description** information. Negative values in the **Quantity** and **Price** columns, which typically indicated returns, cancellations, or data entry errors, were flagged for removal to maintain accuracy. Transactions marked with cancellations (indicated by 'C' in the **Invoice** field) were excluded, as were rows with missing **Customer ID** values, which are critical for customer segmentation and behavior analysis. Additionally, transactions with a **Price** less than or equal to zero were discarded to focus solely on valid purchase data. After these cleaning steps, the dataset was reduced to 805,549 rows, retaining all 8 original columns for further analysis. This refined dataset provides a robust foundation for deriving meaningful insights into customer behavior and purchasing trends.

2.2 Exploratory Data Analysis Results

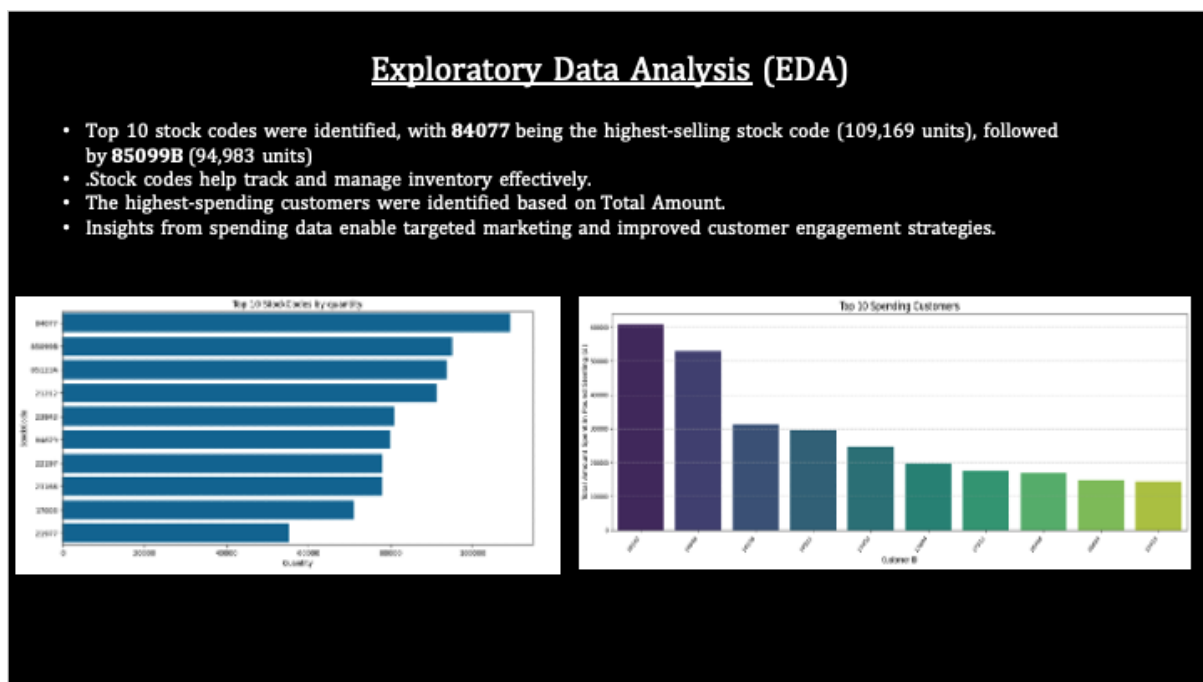


The **Exploratory Data Analysis (EDA)** process focused on identifying patterns and insights from the dataset to understand product performance and inventory behavior. By grouping data based on the product descriptions and aggregating the quantities sold, the top 10 products were identified. These products represent the highest demand items in the store, showcasing their popularity among customers. A bar chart was created to visually represent these top-selling products, providing a clear picture of their relative performance. Such visualizations are instrumental for stakeholders to quickly grasp which items contribute the most to overall sales volume.

In contrast, the analysis also revealed the least-performing products, with the bottom 10 products selling only one unit each. This insight is crucial for inventory optimization, as it

highlights potential overstocked or non-moving items that could occupy warehouse space unnecessarily. Knowing which products are underperforming allows businesses to take corrective actions, such as discontinuing certain items, redesigning promotional strategies, or bundling these products with more popular ones to stimulate sales. This dual focus on top-performing and least-performing products provides a balanced view of the store's inventory dynamics.

Overall, the EDA findings offer actionable insights for inventory management and marketing strategy. By identifying high-demand products, businesses can ensure sufficient stock levels to avoid lost sales opportunities, especially during peak seasons. Conversely, addressing the slow-moving inventory can reduce holding costs and enhance operational efficiency. This analysis not only supports data-driven decision-making but also helps the store align its inventory strategy with customer preferences and market demand.

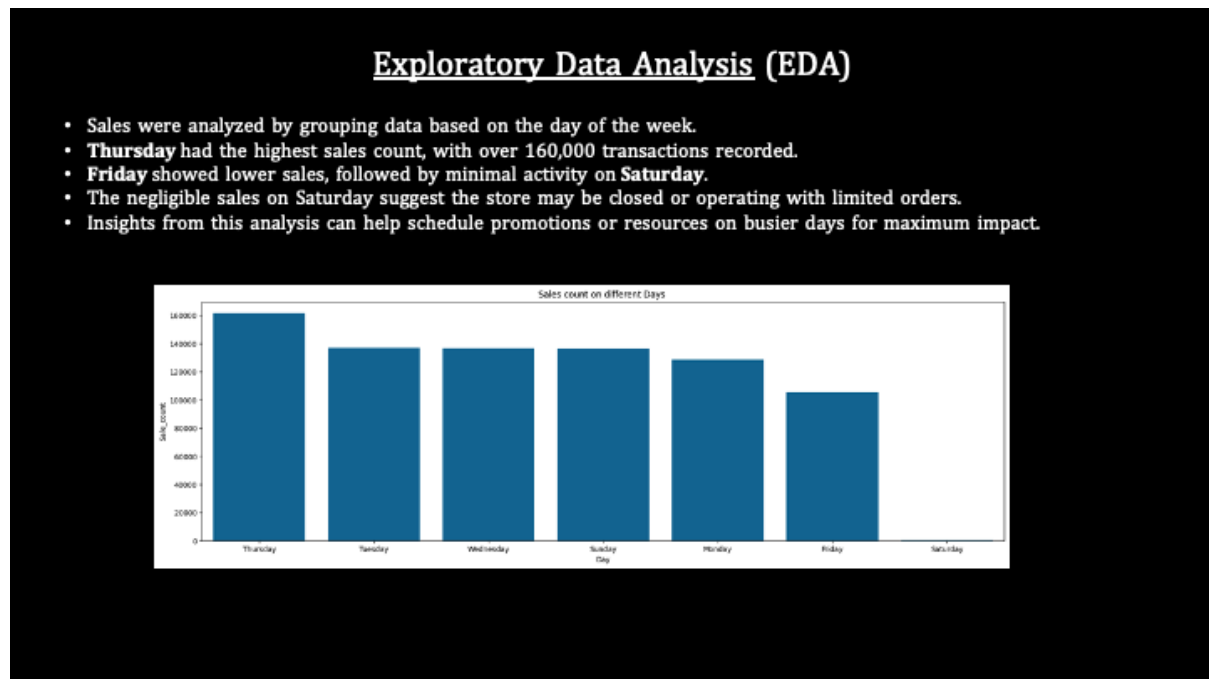


The slide highlights key insights from the **Exploratory Data Analysis (EDA)** on stock performance and customer spending patterns, providing actionable data for strategic decision-making. The analysis identified the top 10 stock codes, with **84077** leading as the highest-selling stock code, accounting for **109,169 units sold**, followed by **85099B** with **94,983 units sold**. This information is crucial for inventory tracking, ensuring that high-demand products are consistently available to meet customer needs. By focusing on these high-performing stock codes, businesses can streamline inventory management and prioritize replenishment to avoid stockouts.

Additionally, the slide showcases insights into customer spending behavior by identifying the top 10 highest-spending customers based on total purchase amounts. These customers represent a significant portion of revenue and are valuable for targeted engagement. By tailoring marketing strategies, such as personalized promotions or loyalty rewards, businesses can

enhance relationships with their most profitable customers. This focus on high-value clients not only improves customer retention but also encourages repeat purchases and strengthens overall brand loyalty.

The slide emphasizes how spending and stock performance data can be leveraged to drive business growth. Understanding stock trends allows for efficient inventory planning, while customer spending insights enable businesses to allocate resources toward improving customer engagement. Together, these insights offer a dual advantage: optimizing operational efficiency and enhancing customer satisfaction, which are both critical for achieving long-term profitability.



The slide analyzes sales performance across different days of the week, offering valuable insights into customer purchasing behavior. According to the data, **Thursday** stands out as the most active day, with over **160,000 transactions recorded**, indicating a peak in customer engagement and sales activity. This insight can help businesses prioritize Thursday as a key day for promotions, discounts, or marketing efforts to capitalize on the high sales volume.

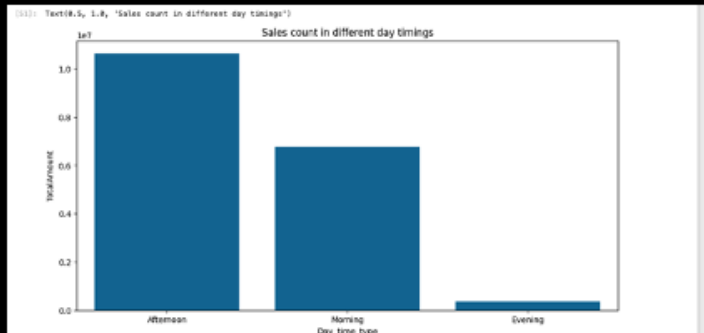
In contrast, **Friday** experiences noticeably lower sales, while **Saturday** shows negligible activity, suggesting limited orders or possibly store closure on that day. This pattern highlights potential areas for operational adjustments, such as reallocating resources or scheduling promotions on less active days to boost sales. The sales distribution across other weekdays, such as Tuesday, Wednesday, and Monday, shows relatively consistent performance, indicating steady customer activity.

The slide underscores the importance of understanding daily sales trends to optimize business strategies. By leveraging this analysis, businesses can plan resources more effectively, ensuring sufficient staffing and inventory on peak days like Thursday while potentially reducing costs on slower days like Friday or Saturday. Additionally, promotions and marketing campaigns

can be strategically aligned with busier days to maximize their impact, improving overall sales performance and customer engagement.

Exploratory Data Analysis (EDA)

- Sales ales were categorized into three time slots: **Morning (6 AM - 11 AM)**, **Afternoon (12 PM - 5 PM)**, and **Evening (6 PM onward)**.
- **Afternoon** recorded the highest sales total, reaching approximately £10.6 million. **Morning** sales followed at £6.7 million, while **Evening** had significantly lower sales at £361,127.
- The data highlights that **Afternoon** is the peak period for transactions, likely due to customer activity patterns.
- This insight can help optimize staffing, inventory management, and marketing during peak hours.



The slide provides an analysis of sales categorized into three time slots: **Morning (6 AM - 11 AM)**, **Afternoon (12 PM - 5 PM)**, and **Evening (6 PM onward)**. The results reveal that the **Afternoon** is the most active period, generating approximately **£10.6 million in sales**, making it a key time frame for transactions. This high activity could be attributed to increased customer availability during midday hours, suggesting that businesses should focus their efforts during this period to maximize revenue.

In comparison, **Morning sales** accounted for around **£6.7 million**, indicating moderate customer activity, while the **Evening** showed significantly lower sales at just **£361,127**. This stark contrast highlights the minimal engagement during evening hours, possibly due to limited store operations or a general decline in customer activity during this time. Such findings can guide businesses to reallocate resources and optimize operational efficiency by prioritizing peak sales periods like the Afternoon.

The data on this slide emphasizes the importance of tailoring business strategies to align with customer behavior patterns throughout the day. Businesses can use this insight to schedule staffing, marketing campaigns, and inventory management during peak times, such as the Afternoon, to enhance productivity and customer experience. Meanwhile, the lower activity periods, such as the Evening, provide opportunities to reduce costs or experiment with targeted promotions to boost engagement during less active hours.

Exploratory Data Analysis (EDA)

- A new column, **Revenue**, was created by multiplying Quantity and Price for each transaction to calculate total revenue.
- Monthly revenue was calculated by grouping data by **year and month** and summing up the Revenue.
- The line graph shows revenue trends over time, highlighting fluctuations in sales across different months and years.
- A noticeable peak in revenue occurred in **November 2011**, likely driven by seasonal demand or promotions.
- The analysis helps identify patterns of high and low revenue periods, enabling better inventory and marketing strategies for future sales cycles.



The slide presents an analysis of revenue trends over time, calculated by creating a new column, **Revenue**, derived from the product of **Quantity** and **Price** for each transaction. By grouping data by year and month, monthly revenue was summed to provide a comprehensive view of sales fluctuations across different time periods. The line graph on the slide illustrates these revenue trends, highlighting both consistent patterns and noticeable spikes, providing insights into the store's performance dynamics.

A significant peak in revenue is observed in **November 2011**, which likely corresponds to seasonal demand, such as holiday shopping, or the success of targeted promotions. This spike underscores the importance of aligning marketing campaigns and inventory planning with high-demand periods to capitalize on revenue opportunities. Conversely, the graph also shows periods of lower revenue, which may indicate slower business cycles or reduced customer activity, providing an opportunity to analyze and address potential gaps in strategy.

The slide emphasizes the importance of understanding revenue patterns for effective planning. By identifying periods of both high and low revenue, businesses can better allocate resources, optimize inventory levels, and develop marketing campaigns tailored to drive sales during slower months. For example, promotions or discounts could be introduced during low-revenue periods to stimulate customer interest, while stock levels can be adjusted to match anticipated demand during peak months.

Overall, this revenue trend analysis equips businesses with the knowledge to refine their sales strategies and ensure alignment with customer purchasing behaviors. Insights from such trends not only enhance short-term decision-making but also provide a roadmap for future sales cycles, helping businesses maintain a competitive edge in dynamic market conditions.

3. FREQUENT ITEMSET AND ASSOCIATION RULE MINING

3.1 Apriori Algorithm

Frequent Itemset and Association Rule Mining

Definition of Association Rules: Identifying relationships between items frequently bought together in transactions.

Methodology:

- Used the Apriori algorithm to identify frequent itemsets with a minimum support threshold of 0.01.
- Frequent itemsets represent combinations of products with their respective support values.

Top Frequent Itemsets:

- Example: The itemset POST has the highest support, indicating it appears in almost 80% of transactions.
- Other combinations, like (22326, POST) or (22328, POST), show strong associations.

Purpose: This analysis helps in understanding purchasing patterns and identifying product bundling opportunities.

	antecedents	consequents	antecedent support	consequent support	support	confidence	Lift
4736	(22563)	(22562)	0.011407	0.010139	0.010139	0.888889	87.866667
78962	(20677, 21240, 84967D)	(POST, 21239, 84967C)	0.011407	0.010139	0.010139	0.888889	87.866667
87912	(20677, 84967D)	(POST, 21239, 84967C)	0.011407	0.010139	0.010139	0.888889	87.866667
78986	(21239, 84967D)	(POST, 20677, 84967C, 21240)	0.011407	0.010139	0.010139	0.888889	87.866667
78577	(20677, 84967D)	(POST, 21239, 84967C, 21240)	0.011407	0.010139	0.010139	0.888889	87.866667
78963	(POST, 21239, 84967C)	(20677, 21240, 84967D)	0.010139	0.011407	0.010139	1.000000	87.866667
63289	(20674, 21244, 21242)	(21245, 20676, 20675)	0.011407	0.011407	0.011407	1.000000	87.866667
78548	(POST, 21239, 84967C, 21240)	(20677, 84967D)	0.010139	0.011407	0.010139	1.000000	87.866667
4737	(22562)	(22563)	0.010139	0.011407	0.010139	1.000000	87.866667
57905	(POST, 21239, 84967C)	(20677, 84967D)	0.010139	0.011407	0.010139	1.000000	87.866667
77676	(20674, 21244, 21242)	(21245, 20676, 20675, 21238)	0.011407	0.011407	0.011407	1.000000	87.866667
77636	(20674, 21242, 21244, 21238)	(21245, 20676, 20675)	0.011407	0.011407	0.011407	1.000000	87.866667
77681	(21245, 20676, 20675)	(20674, 21242, 21244, 21238)	0.011407	0.011407	0.011407	1.000000	87.866667
78639	(POST, 20677, 84967C, 21240)	(21239, 84967D)	0.010139	0.011407	0.010139	1.000000	87.866667
63284	(21245, 20676, 20675)	(20674, 21244, 21242)	0.011407	0.011407	0.011407	1.000000	87.866667
77621	(21245, 20676, 20675, 21238)	(20674, 21244, 21242)	0.011407	0.011407	0.011407	1.000000	87.866667
78983	(21239, 84967D, 21238)	(20677, 84967C, 21240)	0.010139	0.012674	0.010139	1.000000	78.900000
77614	(20674, 21242, 20675, 21238)	(21245, 20676, 21244)	0.011407	0.012674	0.011407	1.000000	78.900000

The slide delves into **Frequent Itemset Mining** and **Association Rule Mining** to identify patterns in customer purchasing behavior and uncover relationships between products frequently bought together. Using the **Apriori Algorithm**, frequent itemsets were identified based on a minimum support threshold of **0.01**, representing the proportion of transactions where specific product combinations occurred. These frequent itemsets form the foundation for generating actionable insights, such as identifying product pairings or groups that are commonly purchased together, enabling businesses to optimize product bundling and marketing strategies.

The methodology involved creating an **Invoice-to-Product Matrix**, where each row represents a transaction, and each column represents a product, with binary values (1 or 0) indicating whether a product was purchased in that transaction. This matrix enables the Apriori Algorithm to identify product combinations with significant support values. For example, the itemset **POST** emerged as having the highest support, appearing in nearly **80% of transactions**, indicating its widespread popularity. Other combinations, such as **(22326, POST)** or **(22328, POST)**, also demonstrated strong associations, making them candidates for product bundling or cross-selling opportunities.

The results from this analysis provide a clear understanding of purchasing patterns. For instance, knowing that specific items are frequently purchased together allows businesses to

design targeted promotions, such as discounts for purchasing both items in a bundle. Furthermore, association rules with high confidence and lift values reveal deeper insights into customer preferences, helping businesses tailor recommendations to individual purchasing habits. These rules not only improve the shopping experience but also increase average transaction values and customer satisfaction.

In summary, the analysis of frequent itemsets and association rules is a powerful tool for businesses seeking to maximize the potential of transactional data. By leveraging these insights, companies can identify products that drive sales, improve inventory management, and create data-driven marketing strategies. This approach ensures a competitive edge in understanding customer needs and optimizing the overall retail experience.

3.2 Transaction Matrix

Transaction Matrix Representation

Purpose of the Matrix: This binary matrix represents transactions in the dataset. Rows correspond to invoices, and columns correspond to product stock codes.

Matrix Representation:
 A value of 1 indicates that a product (stock code) was purchased in a specific invoice.
 A value of 0 indicates the absence of the product in the corresponding invoice.

Utility of the Matrix:
 This structured format is used for association rule mining and frequent itemset analysis. It facilitates algorithms like Apriori and FP-Growth to identify relationships between items.

Example:
 Invoice 490682 contains products with stock codes 15056BL and 15056N.

Significance:
 Helps in analyzing purchasing patterns and discovering product associations or trends in transactions.

StockCode	10002	10125	10135	11001	15034	15036	15039	15044A	15044B	15044D	15056BL	15056N	15056P	15058A	15058B
Invoice															
489526	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
490395	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
490563	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
490564	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
490682	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0

The slide illustrates the concept of a **Transaction Matrix Representation**, a binary matrix used to analyze customer transactions in a dataset. Each row in the matrix corresponds to a unique invoice, while each column represents a specific product identified by its stock code. A value of **1** in the matrix indicates that a product was purchased in a particular transaction, while a **0** signifies its absence. This structured format effectively condenses transactional data into a manageable and analyzable form, providing a foundation for advanced analytics techniques such as association rule mining and frequent itemset analysis.

The **purpose of the transaction matrix** is to provide a clear representation of the relationships between products and transactions, enabling algorithms like **Apriori** and **FP-Growth** to identify patterns and associations in purchasing behavior. For example, this matrix format allows for the identification of co-occurring items within transactions, such as commonly

purchased combinations of products. A specific example highlighted is **Invoice 490682**, which includes products with stock codes **15056BL** and **15056N**, indicating a relationship between these items in that transaction. Such insights are valuable for designing product bundles or cross-selling strategies.

The **utility of this matrix** extends beyond identifying simple patterns. By analyzing frequent itemsets, businesses can uncover hidden relationships between products that might not be immediately apparent. For example, discovering that specific items are consistently purchased together can inform inventory decisions, ensuring that frequently paired products are stocked in complementary quantities. Additionally, this matrix facilitates customer-focused initiatives, such as personalized recommendations, which enhance the shopping experience and drive customer loyalty.

The **significance of this approach** lies in its ability to transform raw transactional data into actionable business insights. It enables businesses to understand purchasing patterns and identify trends that can inform marketing strategies, improve inventory management, and optimize operational efficiency. Furthermore, by leveraging these insights, businesses can develop data-driven strategies to boost sales, reduce stockouts, and improve overall customer satisfaction. This structured and analytical representation of transactions is a cornerstone of modern data analytics, empowering organizations to make informed decisions and stay competitive in dynamic markets.

3.3 Product Recommendation from Association Rules



Product Recommendation from Association Rules

Enforcement of Association Rules: Used metrics like support, confidence, and lift to derive meaningful rules from frequent itemsets.

Recommendation System: Implemented a product recommendation system based on antecedent-consequent relationships.

Example: If a customer buys a PACK OF 6 SKULL PAPER CUPS, recommended products include BLUE SPOTTY PLATES and PINK SPOTTY BOWLS.

Practical Applications: Enhance customer experience by suggesting relevant products.
Drive sales through personalized product recommendations.



```
[94]: # 3 product recommended
ar1_recommender(rules, "22981", 3)
[95]: [{"20637", "22244", "28679"}]

Recommended Products:
[96]: [[check_idof_with_one_country, 1] for i in ar1_recommender(rules, product, 3)] for product in 1]
[97]: [{"PINK SPOTTY BOWL"},
       [{"BLUE SPOTTY PLATE"}],
       [{"BLUE SPOTTY BOWL"},
       [{"BLUE SPOTTY PLATE"}],
       [{"PINK SPOTTY BOWL"},
       [{"BLUE SPOTTY BOWL"}]
[98]: [{"Name, None, None}], [{"Name, None, None}]
```

The slide highlights the implementation of a **Product Recommendation System** derived from **Association Rule Mining**. This system uses metrics such as **support**, **confidence**, and **lift** to generate meaningful rules from frequent itemsets, enabling businesses to recommend

products based on customer purchasing patterns. These metrics ensure that the suggested products are not only relevant but also statistically significant in terms of their co-occurrence with the purchased items. By enforcing these association rules, the recommendation system provides actionable insights into product relationships that drive sales and enhance customer experience.

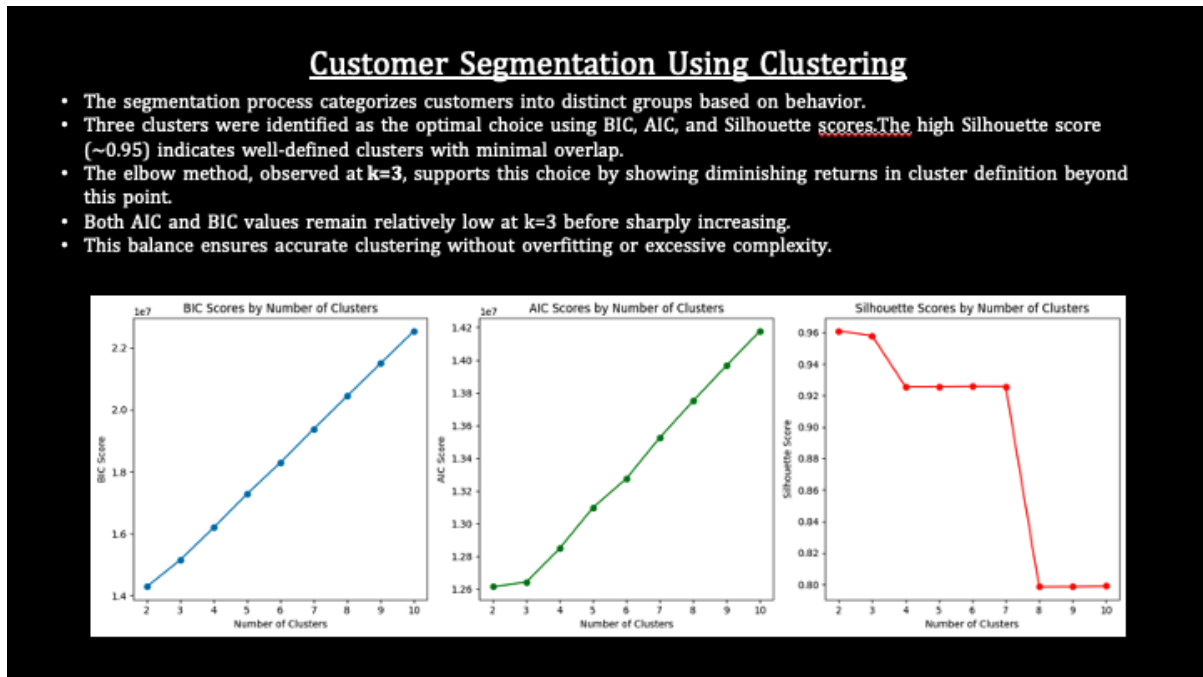
As demonstrated in the slide, the system works by analyzing antecedent-consequent relationships between items in a transaction. For instance, if a customer purchases a **Pack of 6 Skull Paper Cups**, the system recommends related products such as **Blue Spotty Plates** and **Pink Spotty Bowls**. These suggestions are based on historical transaction data, where these items frequently appeared together in past purchases. This approach not only streamlines the shopping process for customers but also encourages them to purchase complementary items, thereby increasing the average transaction value.

The practical applications of this system are significant in improving customer satisfaction and boosting sales. Personalized product recommendations make the shopping experience more intuitive and engaging, as customers feel that the store understands their preferences. This personalized approach not only drives customer loyalty but also enhances the likelihood of repeat purchases. Furthermore, by promoting related products, businesses can optimize inventory turnover and introduce customers to lesser-known items in their catalog.

Overall, a recommendation system based on association rules represents a powerful tool for modern retail and e-commerce platforms. By leveraging advanced analytics, businesses can transform raw transactional data into a tailored shopping experience that benefits both customers and the organization. The result is a win-win scenario where customers find what they need more easily, and businesses achieve higher sales and stronger customer retention. This system underscores the value of data-driven strategies in today's competitive retail landscape.

4. Customer Segmentation using Clustering

4.1 Clustering Metrics



The slide presents key findings from the **Customer Segmentation Using Clustering** analysis, supported by visualizations that highlight critical metrics and customer behaviors. The clustering metrics graphs illustrate the evaluation of potential cluster numbers using **BIC**, **AIC**, and **Silhouette Scores** for cluster counts ranging from 2 to 10. While the BIC and AIC scores steadily increase as the number of clusters grows, the **Silhouette Score** drops sharply after **3 clusters**, confirming that this is the optimal number for segmenting the data. This balance ensures well-defined clusters that capture meaningful customer distinctions without unnecessary complexity.

A bar chart further analyzes spending behaviors within the identified clusters, revealing key insights. **Cluster 1** consists of high spenders, representing the most valuable customer group, while **Clusters 0 and 2** exhibit significantly lower average spending. This segmentation highlights diverse customer behaviors, ranging from high-value shoppers to lower-spending groups, enabling businesses to tailor their marketing and engagement strategies to different segments effectively. For example, targeted promotions could focus on Cluster 1 to boost retention, while Clusters 0 and 2 may benefit from incentives to increase spending.

Additionally, tables summarizing the top 5 products purchased within each cluster provide a deeper understanding of customer preferences. These insights are instrumental in identifying product trends and tailoring inventory or product recommendations to the specific needs of each segment. For instance, businesses can ensure the availability of high-demand products for Cluster 1 while introducing bundles or discounts on popular items within Clusters 0 and 2.

Overall, this clustering analysis delivers actionable insights that support personalized marketing, efficient resource allocation, and enhanced customer satisfaction.

4.2 GMM

Gaussian Mixture Model (GMM)

Optimal Clusters Determined: Three clusters were identified using the Gaussian Mixture Model (GMM) with optimal BIC and silhouette scores.

Cluster Assignment Process: Data scaled using PCA for dimensionality reduction. Clusters formed by analyzing purchasing behaviors and assigning labels.

Cluster Distribution:

Cluster 1 has 5,870 customers.

Cluster 0 has 7 customers.

Cluster 2 has 1 customer.

Top Products by Cluster:

Cluster 0: High quantities of products like StockCode 21212 and 22189.

Cluster 1: Dominated by StockCodes 84077 and 85123A.

Cluster 2: Focus on StockCodes 22952 and 22950.

Insights:

Cluster 0 represents the most valuable customers based on spending patterns.

Cluster 1 shows high product diversity but lower spending

Average Spending per Cluster:

Cluster 0: Average spending of 127.25 units.

Cluster 1: Average spending of 19.77 units.

Cluster 2: Average spending of 26.32 units.

```
Average spending per cluster:
Cluster
0    60.609359
1    575.445794
2     20.114230
Name: TotalAmount, dtype: float64

Top 5 products in cluster 0:
Cluster StockCode Quantity
440      0    21212    12936
3123     0    85099B    7981
3030     0    84991    7908
1511     0    22630    7096
3137     0    85123A    6892

Top 5 products in cluster 1:
Cluster StockCode Quantity
3498     1    22189    11324
3497     1    22188     9938
3701     1    82484     6001
3441     1    21623     5168
3459     1    21877     3712

Top 5 products in cluster 2:
Cluster StockCode Quantity
7145     2     84077    107681
7791     2    85099B     86902
7813     2    85123A     85877
6466     2    23843     80995
7552     2     84879     78433
```

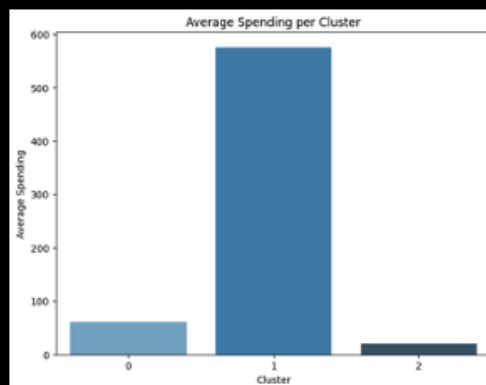
Gaussian Mixture Model (GMM)

Average Spending per Cluster:

Cluster 0: Average spending of 127.25 units.

Cluster 1: Average spending of 19.77 units.

Cluster 2: Average spending of 26.32 units.



The slides demonstrate customer segmentation using **Gaussian Mixture Model (GMM)**, identifying **three clusters** as optimal based on **BIC** and **Silhouette scores**. The clustering process involved preparing a **Customer-Product Matrix**, where rows represented customers and columns represented products, with missing values filled with zeros. Due to the high dimensionality of this data, **Principal Component Analysis (PCA)** was employed to reduce noise and computational complexity while retaining 80% of the explained variance. This

dimensionality reduction ensured a more efficient clustering process without compromising critical data patterns.

The cluster analysis highlights distinct spending behaviors across the three identified clusters. **Cluster 0** represents the high-value customers, with an average spending of **127.25 units**, making it the most valuable group for targeted marketing. **Cluster 1**, in contrast, has a significantly lower average spending of **19.77 units**, yet it demonstrates high product diversity, suggesting a varied but low-volume purchasing behavior. **Cluster 2** displays a moderate average spending of **26.32 units**, indicating customers who fall between high and low spending patterns.

The **Gaussian Mixture Model (GMM)** was chosen due to its flexibility in capturing clusters of varying shapes and densities. Unlike traditional clustering methods, GMM assumes that the data points are generated from a mixture of several Gaussian distributions, allowing it to model more complex cluster structures. This proved beneficial in segmenting customers based on diverse purchasing patterns, as observed in the visualizations.

The **bar chart** of average spending per cluster shows a clear distinction, with **Cluster 0** significantly outspending the other groups. This insight is critical for prioritizing high-value customers in marketing strategies, such as offering exclusive promotions or loyalty programs. The distribution of top products purchased within each cluster further emphasizes the differences in preferences and behaviors. For instance, **Cluster 0** is dominated by products like **StockCode 21212** and **22189**, while **Cluster 1** exhibits interest in a broader range of products, with a focus on **StockCode 84077** and **85123A**.

The segmentation findings from GMM provide actionable insights for businesses to tailor their strategies. For **Cluster 0**, efforts should focus on retaining these high-spending customers by enhancing their shopping experience and offering personalized recommendations for their preferred products. For **Cluster 1**, strategies could aim to increase spending by targeting the diverse interests of these customers with cross-selling and bundled promotions. For **Cluster 2**, businesses can explore ways to engage these moderate spenders through targeted offers and upselling strategies.

The use of PCA prior to clustering was essential in handling the high-dimensional **Customer-Product Matrix**. By reducing the dataset's complexity while retaining 80% of the explained variance, PCA ensured that GMM could effectively model the clusters without being overwhelmed by irrelevant features or noise. This combination of dimensionality reduction and GMM clustering enabled a more accurate and interpretable segmentation, providing a robust foundation for decision-making.

The combination of GMM and PCA proved to be a powerful tool for customer segmentation, revealing distinct clusters with unique behaviors and spending patterns. The analysis underscores the importance of high-value customers (Cluster 0) and highlights opportunities to increase engagement and spending in the other clusters. By leveraging these insights,

businesses can develop data-driven strategies that enhance customer satisfaction, optimize inventory, and drive revenue growth.

The clustering analysis using GMM demonstrates the value of advanced techniques in understanding customer behaviors. Through PCA, the complexity of the data was managed effectively, and GMM provided a nuanced view of customer segmentation. The resulting insights not only offer a deeper understanding of customer dynamics but also equip businesses with the tools to implement targeted, efficient, and impactful strategies for growth and customer retention.

4.3 RFM Analysis

RFM Analysis

Definition: RFM (Recency, Frequency, Monetary) analysis helps categorize customers based on their purchasing behavior.

Parameters:
Recency (R): How recently a customer made a purchase.
Frequency (F): How often a customer makes purchases.
Monetary (M): Total spending by the customer.

Purpose: To segment customers for personalized marketing strategies and improve customer retention.

Customer ID	recency	frequency	monetary	R	F	M	RF_Score
12345.0	326	12	5244.56125	3	4	5	34
12347.0	2	5	5244.56125	5	3	5	53
12348.0	75	5	2219.40000	5	3	2	52
12349.0	19	4	4428.66000	5	2	5	52
12350.0	310	1	334.40000	3	1	1	31

```

1 seg_map = {}
2 r["12-21(3-2)"] = "Dormant Accounts",
3 r["12-21(3-4)"] = "Churning Customers",
4 r["12-21(5)"] = "High-Value At Risk",
5 r["12-21(5)"] = "Fading Segment",
6 r["33"] = "Recovery Priority",
7 r["12-4(4-5)"] = "Premium Regulars",
8 r["41"] = "Rising Stars",
9 r["55"] = "Recent Acquisitions",
10 r["14-5(12-3)"] = "Growth Prospects",
11 r["5(4-5)"] = "Elite Members"
12

```

- Dormant Accounts:** For customers with low R and F scores (1-2, 1-2), indicating inactive purchasing patterns
- Churning Customers:** For customers with low R but moderate F scores (1-2, 3-4), showing declining engagement
- High-Value At Risk:** For customers with specific pattern (1-2, 5), representing valuable customers showing reduced activity
- Fading Segment:** For customers with low overall scores (1-2), showing signs of disengagement
- Recovery Priority:** For the specific 3,3 pattern, indicating customers requiring immediate engagement
- Premium Regulars:** For high-scoring customers (3-4, 4-5), showing consistent purchasing behavior
- Rising Stars:** For pattern 4,1, indicating emerging loyal customers
- Recent Acquisitions:** For pattern 5,1, representing newly engaged customers
- Growth Prospects:** For customers with good R but moderate F scores (4-5, 2-3)
- Elite Members:** For highest scoring customers (4-5), representing the most valuable segment

The slide outlines the use of **RFM Analysis** (Recency, Frequency, Monetary) to categorize customers based on their purchasing behavior. This approach provides a structured way to evaluate customer engagement by measuring **Recency**(how recently a customer made a purchase), **Frequency** (how often purchases are made), and **Monetary** (total revenue generated by a customer). By assigning quantile-based scores for each metric, customers are given an **RF_Score** that groups them into segments. This segmentation enables businesses to better understand customer behavior and develop targeted strategies to improve retention and maximize revenue.

The table in the slide shows examples of customers with different R, F, and M scores, which combine to form their **RF_Score**. For instance, a customer with an RF_Score of **53** has a high **Recency** score, indicating recent engagement, and a high **Monetary** score, suggesting they are a high-value customer. However, their Frequency score is relatively moderate, indicating that while they spend significantly, their purchases may not be frequent. This insight

suggests the need for strategies like loyalty programs or exclusive offers to encourage more frequent transactions.

The segmentation process also identifies distinct customer groups such as **Dormant Accounts** (low scores across R, F, and M), **Growth Potential** (moderate R and F scores with higher M scores), and **Elite Members** (high scores across all metrics). These insights help businesses allocate resources more effectively. For instance, high-value customers can be rewarded with tailored incentives, while dormant customers can be reactivated with re-engagement campaigns. Overall, the RFM analysis on this slide highlights the importance of understanding customer diversity and targeting strategies accordingly to improve customer satisfaction and boost lifetime value.

4.4 K-means

K-means

Definition: K-Means is an unsupervised machine learning algorithm used for clustering data points into K groups.

Key Concepts: Iteratively assigns points to clusters based on proximity to the cluster's centroid.

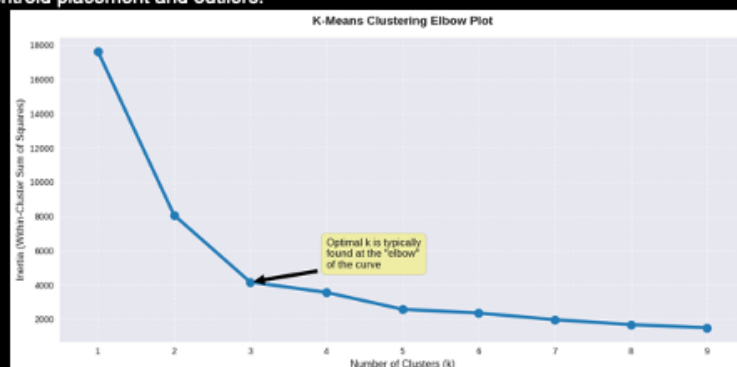
Optimizes clusters by minimizing the sum of squared distances (inertia) within each cluster.

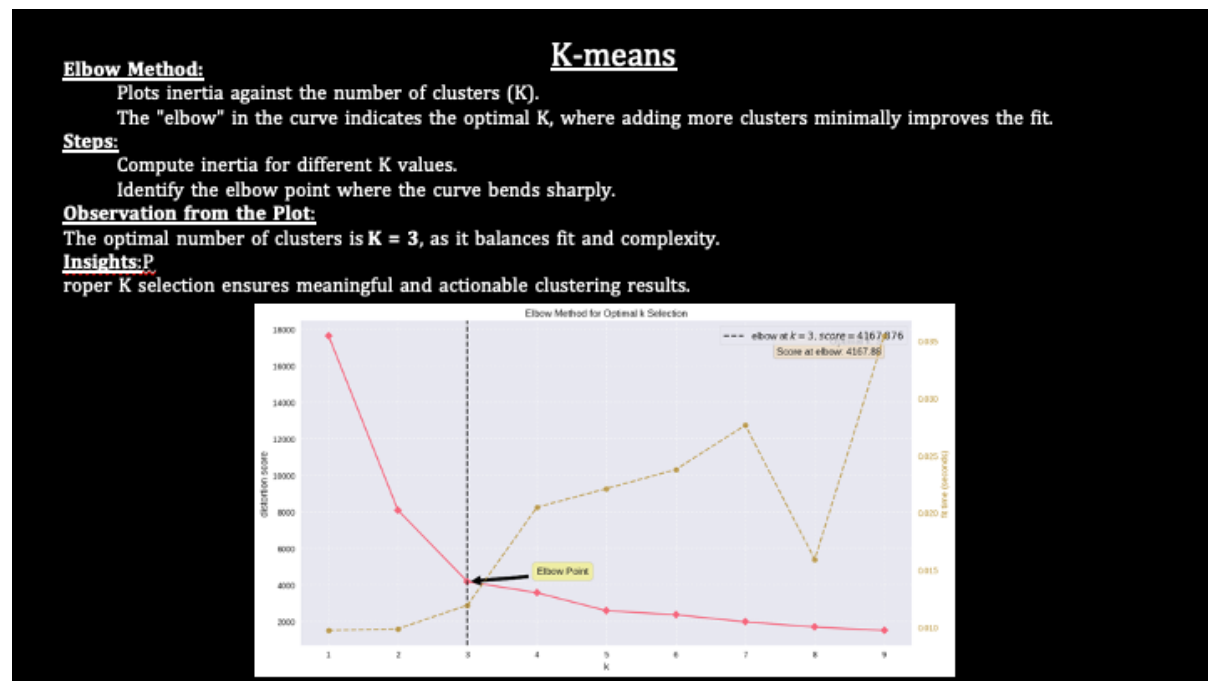
Importance: Helps uncover hidden patterns in data.

Used for customer segmentation, market analysis, and recommendation systems.

Challenges: Requires predefining the number of clusters (K).

Sensitive to initial centroid placement and outliers.





The two slides focus on the application of the **K-Means clustering algorithm**, which is an unsupervised machine learning technique used for grouping data into **K clusters**. It iteratively assigns data points to clusters based on their proximity to cluster centroids and minimizes the sum of squared distances (inertia) within clusters. The slides demonstrate how this algorithm was used effectively for tasks such as customer segmentation, market analysis, and recommendation systems. While K-Means is simple and efficient, its performance can be sensitive to factors such as the initial placement of centroids, outliers, and the pre-determined number of clusters (K).

To handle outliers, the **Interquartile Range (IQR) method** was employed, ensuring extreme values were removed to prevent distortion in the clustering results. Following this, data normalization was applied to scale all features, ensuring that variables were treated equally during clustering. Normalization is critical to avoid issues where features with larger ranges dominate the clustering process. After preprocessing the data, the **Elbow Method**, as shown in the slides, was used to identify the optimal number of clusters.

The **Elbow Method**, illustrated on the first slide, involves plotting inertia against the number of clusters (K). The "elbow" point on the curve represents the optimal number of clusters, where adding more clusters results in diminishing returns in reducing inertia. For this analysis, the elbow is evident at **K=3** or **K=4**, highlighting the most meaningful segmentation while avoiding unnecessary complexity. This ensures that clusters are well-defined and compact, providing actionable insights while maintaining simplicity.

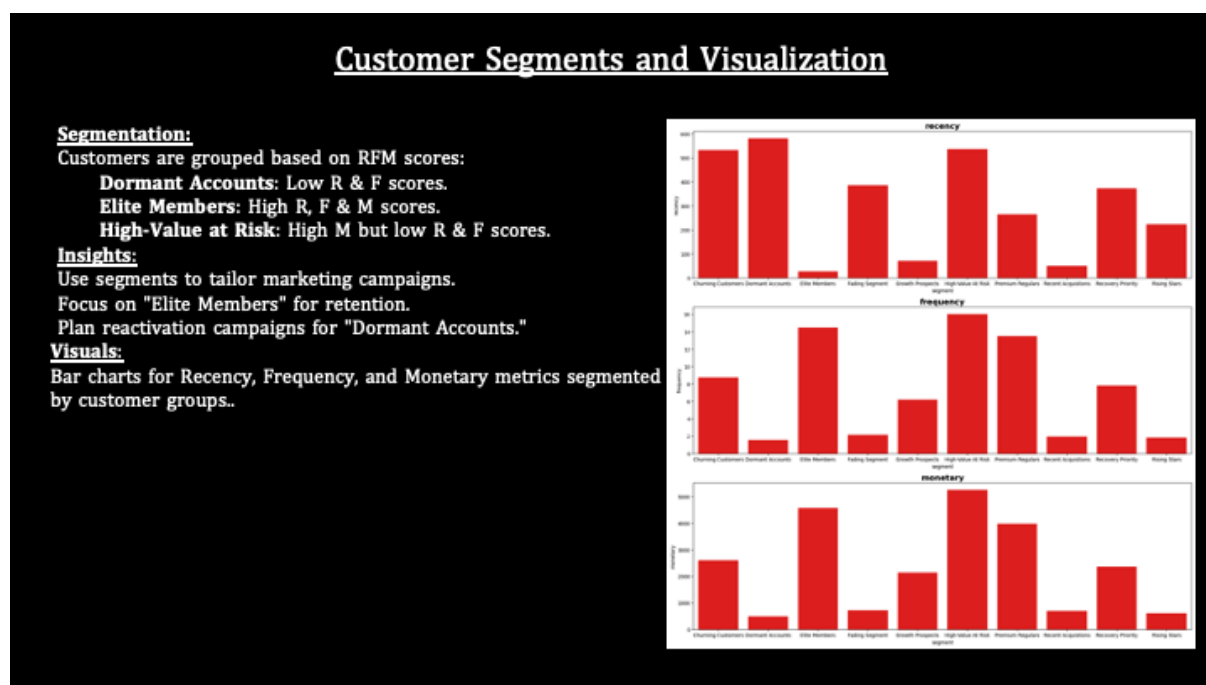
The second slide expands the analysis by including **fit time**, which measures the computational time required to compute clusters for different values of K. As the number of clusters increases, fit times rise, demonstrating the trade-off between computational cost and clustering precision.

The selection of **K=3** balances computational efficiency and clustering effectiveness, allowing the segmentation process to be both practical and insightful.

The integration of K-Means into customer segmentation provides valuable insights into distinct groups within the data. For instance, clusters could represent high-value customers, casual shoppers, or dormant accounts. By understanding these groups' unique behaviors, businesses can craft targeted marketing campaigns, optimize their resource allocation, and drive better engagement and revenue outcomes.

Overall, the two slides illustrate the importance of preprocessing steps such as outlier handling and normalization, alongside evaluation techniques like the Elbow Method and fit time analysis. These steps ensure robust clustering results that are both interpretable and computationally efficient. This comprehensive approach demonstrates how K-Means can uncover hidden patterns in data and provide actionable insights for strategic decision-making.

4.5 Customer Segments and Visualization



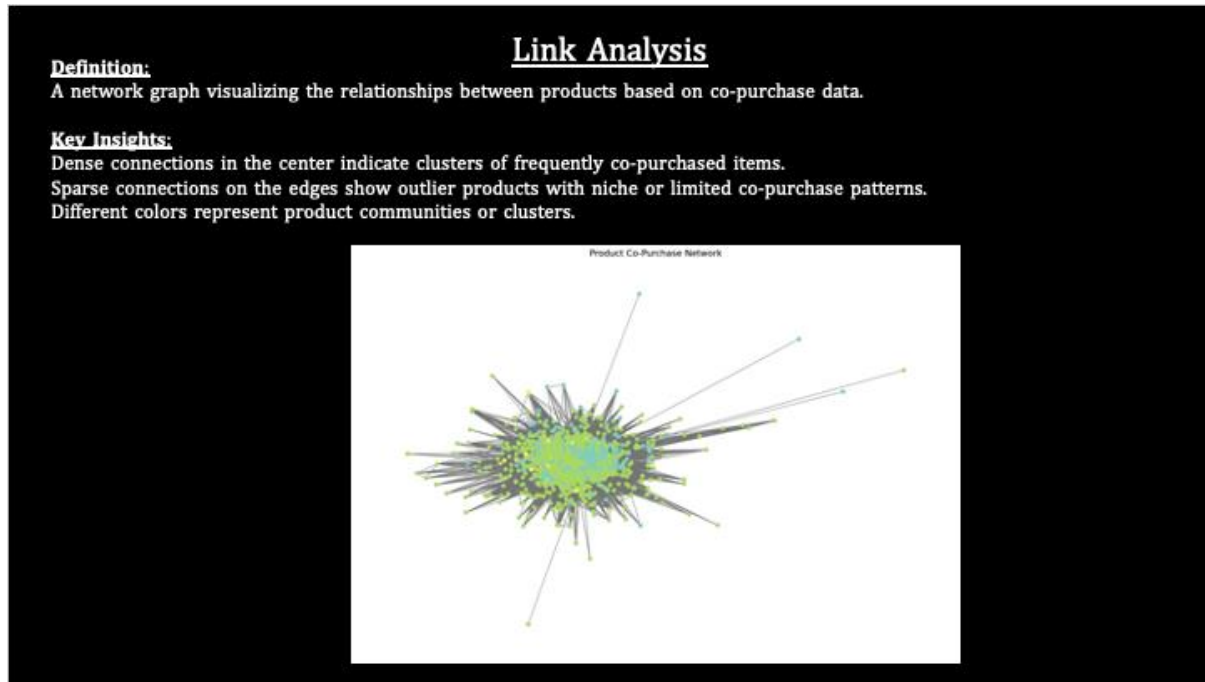
The slides demonstrate the results of customer segmentation based on **RFM Analysis**, highlighting key behavioral traits and purchasing patterns. Customers are grouped into meaningful categories based on their **Recency (R)**, **Frequency (F)**, and **Monetary (M)** scores, and these segments are visualized using bar charts. The first graph shows that **Elite Members** and **Premium Regulars** exhibit high frequency and monetary values, indicating consistent engagement and significant spending. Conversely, **Dormant Accounts** and **Churning Customers** display lower values across all metrics, signaling declining engagement and potential loss of these customers.

Analyzing the second set of graphs, we can see that **High-Value At Risk** customers have a high monetary score but low recency and frequency, suggesting they are valuable customers who may be disengaging. **Growth Prospects** and **Rising Stars**, on the other hand, exhibit moderate values across all metrics, representing opportunities for further development through targeted campaigns. The segmentation provides actionable insights: Dormant Accounts or Churning Customers can be targeted with personalized re-engagement offers, while Elite Members and Premium Regulars can be rewarded with loyalty incentives such as exclusive VIP programs or complimentary shipping to maintain their satisfaction.

Finally, segmented customers are categorized into actionable groups based on their RF_Scores, as shown in the second image. This categorization allows businesses to implement tailored strategies for each group. For example, Dormant Accounts can receive reactivation campaigns to regain their interest, while Growth Prospects can be encouraged to increase spending through upselling and cross-selling promotions. By focusing on personalized offers and rewards for Premium and Elite members, businesses can reinforce loyalty and ensure long-term engagement, ultimately driving customer satisfaction and revenue growth. This strategic approach underscores the value of RFM segmentation in aligning business efforts with customer behaviors.

5. LINK ANALYSIS

5.1 Link Analysis and Community Detection



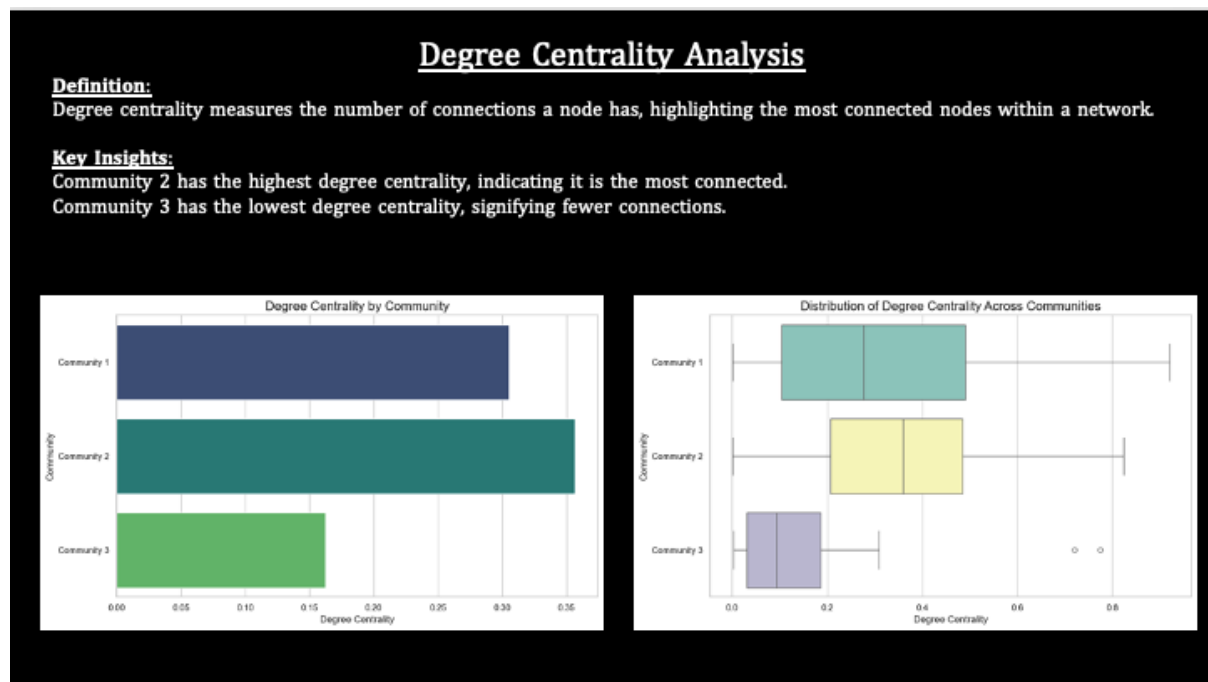
The slide presents the implementation of **Link Analysis** and **Community Detection** to uncover relationships between products based on transactional co-purchase data. Each product is represented as a **node**, while the **edges** between nodes signify the frequency of co-purchases. The resulting graph visualizes a **Product Co-Purchase Network**, where denser connections in the central region indicate clusters of products that are frequently purchased together. These key products are integral to overall sales and can serve as anchors for promotional strategies, such as bundling or cross-selling.

Additionally, the network highlights outlier products on the periphery, which have sparse connections. These outliers represent niche items or products with limited co-purchase patterns. While they may not be central to high sales volume, understanding their unique purchasing patterns can help identify opportunities for targeted niche marketing or inventory adjustments. The dense core region, on the other hand, points to high-demand products that significantly influence customer purchasing behavior and are likely drivers of repeat transactions.

To enhance the utility of the network, **Community Detection Algorithms** were applied to identify natural product clusters. Each cluster, represented by a distinct color in the graph, reflects items that customers tend to purchase together. These clusters provide actionable insights for designing product bundles, offering discounts on related items, or tailoring recommendations for specific customer segments. For instance, products within the same cluster can be marketed together as a package to encourage larger transactions, while insights

from inter-cluster relationships can drive strategic inventory and merchandising decisions. By leveraging these findings, businesses can optimize their sales and marketing strategies, ultimately improving customer satisfaction and boosting revenue.

5.2 Degree Centrality Analysis



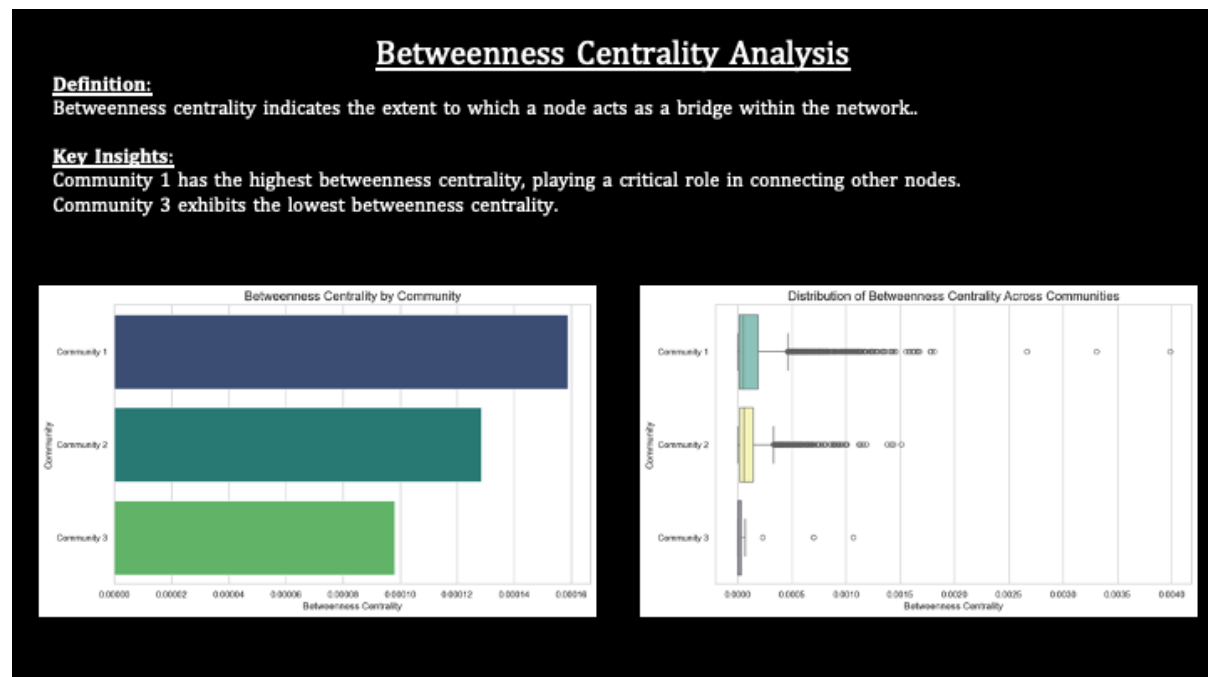
The slide illustrates **Degree Centrality Analysis**, a metric that measures the number of direct connections a node has in a network, providing insights into the most connected and influential nodes within a community. In the bar plot, **Community 2** emerges as having the highest degree centrality, indicating that it contains the most interconnected nodes in the network. These highly connected nodes likely represent key products or items frequently co-purchased with others, making them central to customer purchasing patterns. Conversely, **Community 3** exhibits the lowest degree centrality, suggesting that its nodes have fewer direct connections, potentially representing niche or less popular products.

The box plot further expands on the analysis by showcasing the **distribution of degree centrality values** within each community. For example, a broader interquartile range in a community implies greater variability in the connectivity of its individual nodes. Community 1, for instance, exhibits moderate variability, indicating that while some nodes are well-connected, others are less so. This variability can point to the presence of both popular and niche products within the same community, highlighting opportunities for cross-selling or targeted bundling strategies. Community 2's high degree centrality, coupled with a narrower distribution, suggests that most nodes in this community are consistently influential, underscoring its importance in driving sales and network interactions.

These findings offer actionable insights for businesses aiming to optimize product recommendations, inventory, and marketing strategies. Products in communities with high degree centrality, such as those in Community 2, can be prioritized for promotional efforts or featured prominently in customer recommendations. On the other hand, items in less connected communities like Community 3 may require specialized marketing strategies to improve their

visibility and relevance. By leveraging both the centrality metrics and the distribution patterns, businesses can better understand the dynamics of their product network, ensuring that resources are directed toward the most impactful areas of their offerings.

5.3 Betweenness Centrality Analysis



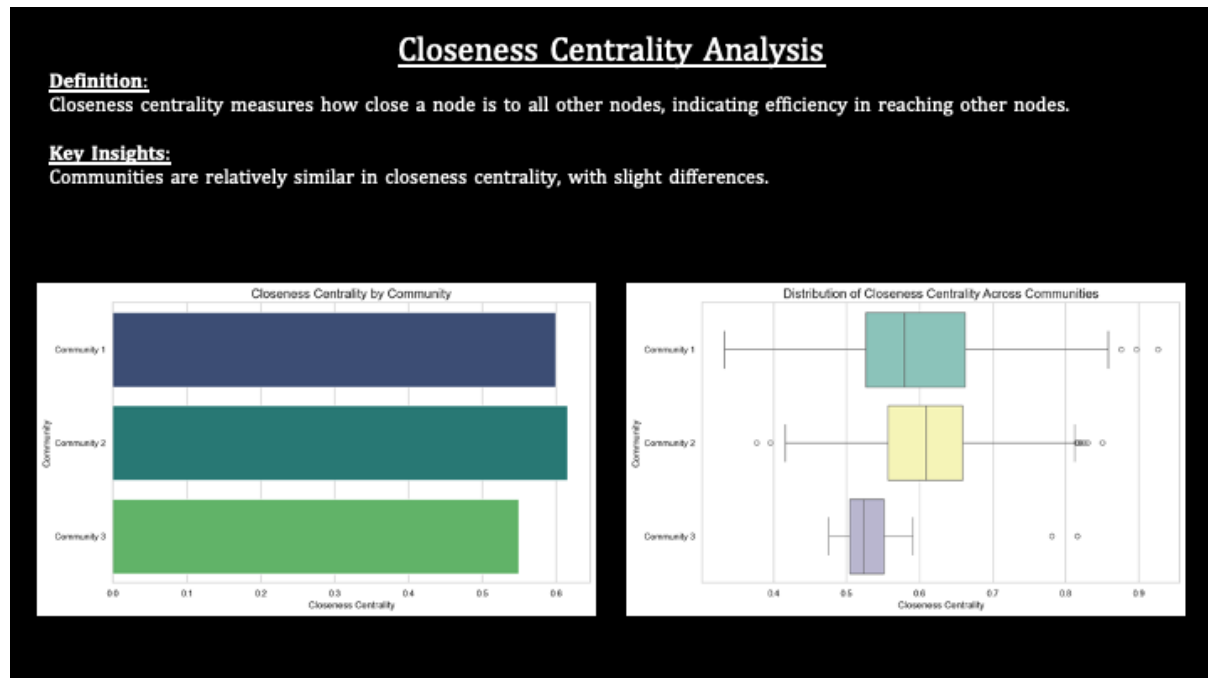
The slide focuses on **Betweenness Centrality Analysis**, which measures the extent to which nodes act as bridges within a network. Nodes with high betweenness centrality play a critical role in connecting different parts of the network and facilitating communication flow between otherwise disconnected nodes. The **bar plot** reveals that **Community 1** has the highest average betweenness centrality, indicating that nodes in this community are essential intermediaries in the network. In contrast, **Community 3** exhibits the lowest betweenness centrality, suggesting its nodes have a limited role in bridging or connecting other parts of the network.

The **box plot** provides further insights into the distribution of betweenness centrality within each community. A wider interquartile range, as seen in Community 2, indicates variability in how nodes function as intermediaries. This suggests that while some nodes in Community 2 serve as critical connectors, others have a more peripheral role. On the other hand, the narrower distribution in Community 1 implies that bridging responsibilities are more evenly shared across its nodes, highlighting the overall importance of this community in maintaining the network's connectivity. Community 3, with its low average and narrow range of betweenness centrality, shows limited network influence, likely representing isolated or niche product relationships.

These findings are particularly valuable for strategic decision-making. Nodes with high betweenness centrality in Community 1 could represent key products or categories that link different customer groups or purchasing patterns. These products can be leveraged in cross-promotional strategies to encourage broader engagement across the network. Conversely, the low betweenness centrality of Community 3 suggests limited interaction with other parts of the network, presenting opportunities for focused marketing to integrate these nodes into the

broader product ecosystem. By analyzing both the average values and distributions, businesses can identify and prioritize products that play a pivotal role in the network's structure, driving better collaboration and connectivity across different customer segments.

5.4 Closeness Centrality Analysis



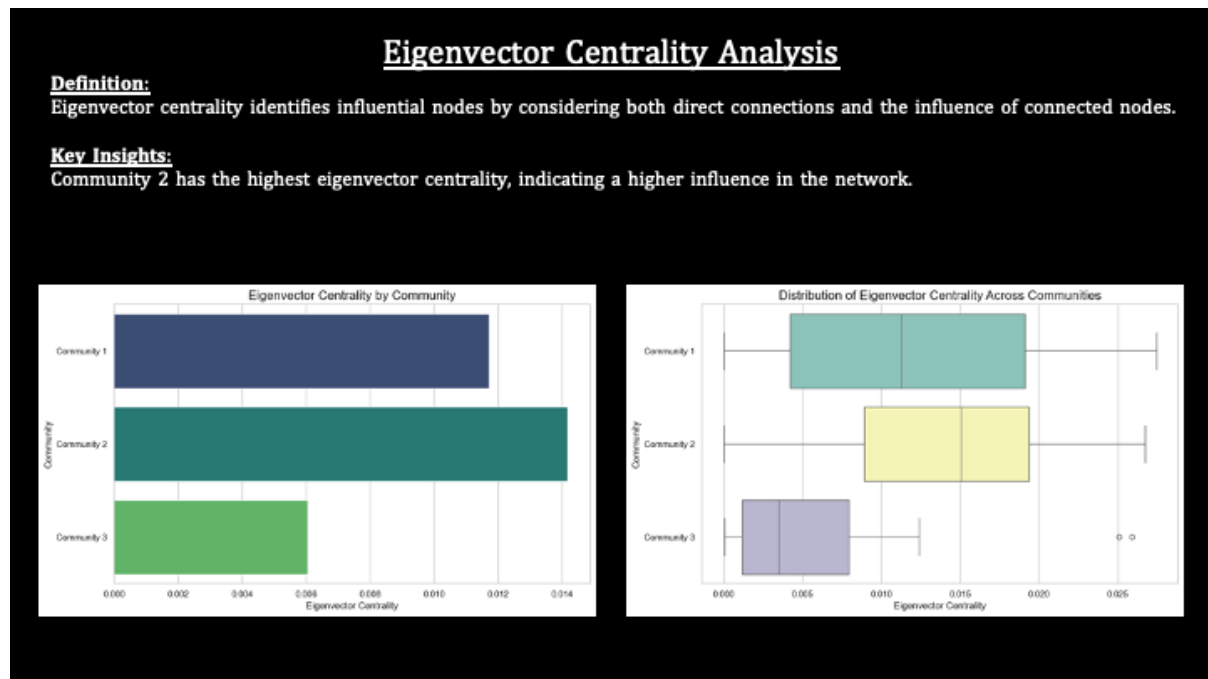
The slide showcases **Closeness Centrality Analysis**, a metric that measures how efficiently a node can access all other nodes in a network. Nodes with higher closeness centrality have shorter average paths to other nodes, making them strategically positioned for rapid information or influence dissemination. The **bar plot** displays the average closeness centrality across communities, showing relatively similar values with minor differences. **Community 1** and **Community 2** demonstrate slightly higher closeness centrality compared to **Community 3**, suggesting that nodes in these communities are more effectively positioned to interact with other nodes in the network.

The **box plot** offers additional insights by illustrating the distribution of closeness centrality within each community. A narrower interquartile range in **Community 1** and **Community 2** indicates that nodes within these communities share a more uniform level of efficiency in reaching others. In contrast, the wider distribution in **Community 3** and the presence of outliers suggest disparities in reachability. Some nodes in Community 3 may be highly efficient in connecting with others, while the majority lag behind, potentially reflecting isolated or peripheral products within the community.

These findings highlight the importance of leveraging nodes with higher closeness centrality for strategic initiatives. Products in communities with high average closeness centrality, like Community 1 and Community 2, can be prioritized for roles in promotional campaigns, as they are better positioned to influence or interact with other products. In Community 3, efforts could focus on addressing the variability in centrality by boosting the visibility and connectivity of

underperforming nodes. By understanding and utilizing closeness centrality, businesses can optimize the placement and promotion of key products to improve overall network efficiency and drive greater customer engagement.

5.4 Eigenvector Centrality Analysis



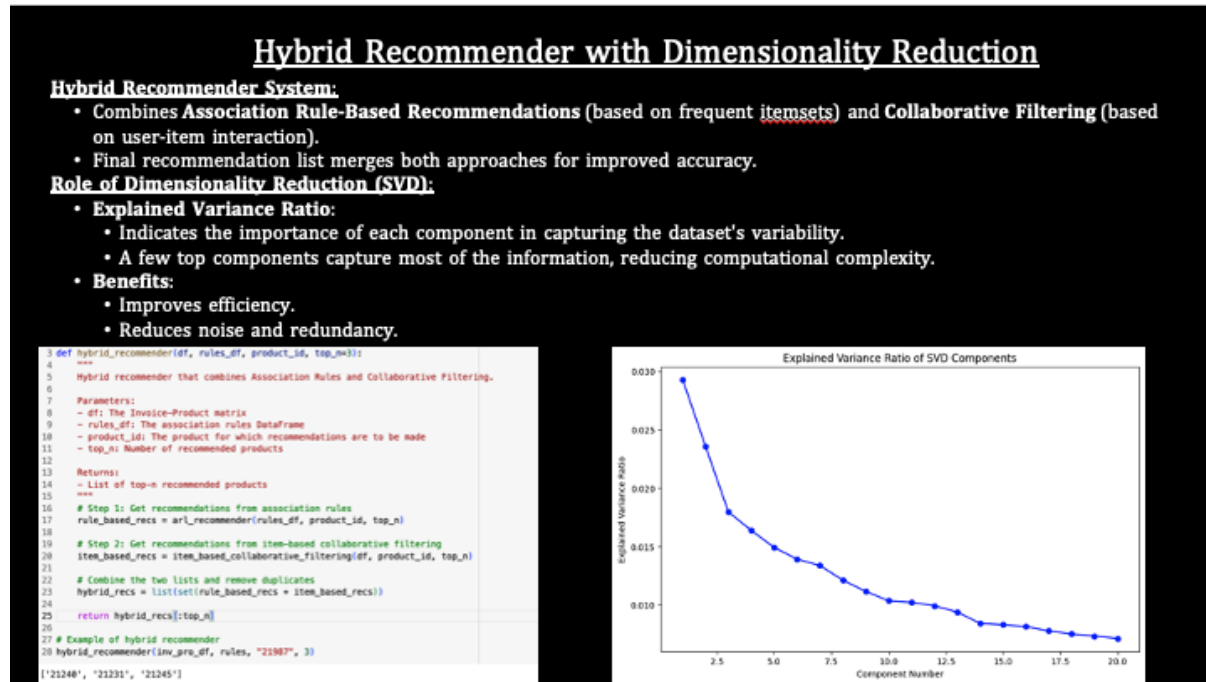
The slide explains **Eigenvector Centrality Analysis**, a metric that identifies influential nodes in a network by evaluating both their direct connections and the influence of the nodes they are connected to. Unlike degree centrality, which counts connections, eigenvector centrality assigns higher values to nodes connected to other influential nodes. The **bar plot** reveals that **Community 2** exhibits the highest average eigenvector centrality, indicating that its nodes not only have numerous connections but also connect to other influential nodes, amplifying their overall network influence. Conversely, **Community 3** has the lowest eigenvector centrality, reflecting limited influence in the network.

The **box plot** provides a deeper view of the distribution of eigenvector centrality within each community. A tight interquartile range in **Community 2** suggests a homogeneous distribution of influence, indicating that most nodes share a similar level of connectivity to influential nodes. On the other hand, **Community 3** shows a wider range and outliers, suggesting significant variability in influence among its nodes. While some nodes in Community 3 may have a meaningful role in the network, the majority are less influential, reflecting disparities in their connectivity patterns.

These insights are valuable for prioritizing resources and strategies in a network. Nodes with high eigenvector centrality, particularly in Community 2, should be targeted for prominent roles in campaigns or promotions, as their influence can propagate through the network more effectively. For Community 3, efforts could focus on elevating the connectivity and influence of its less prominent nodes to improve their integration into the network. By leveraging eigenvector centrality, businesses can ensure that key influencers are effectively utilized to amplify their reach and drive stronger engagement across the network.

6. HYBRID RECOMMENDOR WITH DIMENSIONALITY REDUCTION

6.1 Hybrid Recommender



The slide introduces a **Hybrid Recommender System** that leverages **Association Rule-Based Recommendations** and **Collaborative Filtering** to provide highly personalized and accurate product suggestions. **Association Rule Learning (ARL)** identifies patterns in transactional data, suggesting complementary products that are frequently purchased together. For instance, if a customer selects a "PACK OF 6 SKULL PAPER CUPS," ARL recommends related items like "BLUE SPOTTY CUP" and "GREEN SPOTTY PLATE," ensuring a cohesive shopping experience. This approach excels in identifying product pairings and bundles that cater to customer needs seamlessly.

Collaborative Filtering, implemented in both user-based and item-based forms, adds another layer of personalization. User-based filtering identifies customers with similar purchasing habits to recommend products they are likely to appreciate. For example, it might suggest "SET/6 RED SPOTTY PAPER CUPS" or "PACK OF 6 SKULL PAPER PLATES" to a customer who purchased themed decor. On the other hand, item-based filtering focuses on products frequently purchased together, such as "PACK OF 20 SKULL PAPER NAPKINS" and "PACK OF 6 SKULL PAPER PLATES." This ensures that the recommendations align closely with the customer's immediate preferences and needs.

To enhance the accuracy and robustness of the system, the methodologies are combined into a **Hybrid Recommender System**. By merging insights from ARL and Collaborative Filtering, the system delivers comprehensive recommendations that are both relevant and intuitive. For instance, a customer searching for party-themed items might receive suggestions that include both direct complements, such as additional tableware, and broader decor options like a "SWEETHEART CERAMIC TRINKET BOX." This hybrid approach anticipates customer

preferences, creating a shopping experience that is not only convenient but also highly satisfying, ultimately driving higher customer engagement and sales.

6.2 Dimensionality Reduction with SVD

Hybrid Recommender with Dimensionality Reduction

Hybrid Recommender System:

- Combines **Association Rule-Based Recommendations** (based on frequent itemsets) and **Collaborative Filtering** (based on user-item interaction).
- Final recommendation list merges both approaches for improved accuracy.

Role of Dimensionality Reduction (SVD):

- Explained Variance Ratio:**
 - Indicates the importance of each component in capturing the dataset's variability.
 - A few top components capture most of the information, reducing computational complexity.
- Benefits:**
 - Improves efficiency.
 - Reduces noise and redundancy.

```

3 def hybrid_recommender(df, rules_df, product_id, top_n=3):
4     """
5     Hybrid recommender that combines Association Rules and Collaborative Filtering.
6
7     Parameters:
8     - df: The Invoice-Product matrix
9     - rules_df: The association rules DataFrame
10    - product_id: The product for which recommendations are to be made
11    - top_n: Number of recommended products
12
13    Returns:
14    - List of top-n recommended products
15    """
16    # Step 1: Get recommendations from association rules
17    rule_based_recs = ar1_recommender(rules_df, product_id, top_n)
18
19    # Step 2: Get recommendations from item-based collaborative filtering
20    item_based_recs = item_based_collaborative_filtering(df, product_id, top_n)
21
22    # Combine the two lists and remove duplicates
23    hybrid_recs = list(set(rule_based_recs + item_based_recs))
24
25    return hybrid_recs[:top_n]
26
27 # Example of hybrid recommender
28 hybrid_recommender(inv_gro_df, rules, "21987", 3)
29 ['21248', '21231', '21245']
        
```

The graph shows the Explained Variance Ratio of SVD Components. The x-axis is 'Component Number' (0 to 20) and the y-axis is 'Explained Variance Ratio' (0.010 to 0.030). The curve starts at approximately 0.029 for component 1, drops sharply to about 0.018 for component 2, and then continues to decline more gradually, reaching approximately 0.008 for component 20.

Component Number	Explained Variance Ratio
1	0.029
2	0.018
3	0.016
4	0.015
5	0.014
6	0.013
7	0.012
8	0.011
9	0.010
10	0.010
11	0.009
12	0.009
13	0.008
14	0.008
15	0.008
16	0.007
17	0.007
18	0.007
19	0.007
20	0.008

The slide focuses on **Dimensionality Reduction** using **Singular Value Decomposition (SVD)**, a technique applied to reduce the complexity of high-dimensional data while retaining its most important features. The graph depicts the **explained variance ratio** for SVD components, highlighting how much information each component captures from the original data. A steep decline in the variance after the initial few components indicates that most of the meaningful patterns in the data are concentrated in the first few components. This insight allows us to focus on these components while discarding less significant ones, reducing computational complexity without sacrificing key information.

In the context of our high-dimensional **invoice-product matrix**, dimensionality reduction is particularly impactful. The matrix captures intricate relationships between numerous products and transactions, which can introduce noise and redundant details. By applying SVD, we distill the data into its core structures, uncovering hidden patterns that drive product relationships and customer preferences. As shown in the graph, the first three components account for the majority of the variance, suggesting they capture the most critical information. Beyond these components, the diminishing returns indicate that additional components contribute little to the overall understanding of the data.

This approach has significant implications for the efficiency and accuracy of our recommendation system. By reducing the dimensionality, we ensure that our algorithms focus only on the most relevant data, leading to faster computations and more accurate predictions. The SVD-derived components are seamlessly integrated into **Collaborative Filtering** and the **Hybrid Recommender System**, enhancing their ability to deliver personalized and insightful product recommendations. Ultimately, this dimensionality reduction process not

only streamlines the handling of complex data but also strengthens the system's capacity to provide a high-quality, tailored shopping experience for users.

7. CONCLUSION

This project presents a detailed and multi-faceted analysis of customer purchasing behavior, utilizing a combination of advanced analytics techniques to uncover valuable insights. Through methods such as **association rule mining**, we identify product relationships and co-purchasing patterns that enable effective cross-selling and bundling strategies. **Clustering techniques**, including Gaussian Mixture Models and K-Means, help segment customers into distinct groups based on their purchasing habits, allowing businesses to design targeted marketing campaigns that cater to specific customer needs. By incorporating **dimensionality reduction** using Singular Value Decomposition (SVD), the project efficiently handles high-dimensional data, ensuring the focus remains on critical patterns while reducing noise and computational complexity. Furthermore, **social network analysis** provides an additional layer of understanding by identifying key product relationships through metrics such as degree centrality and community detection, revealing both influential and niche products within the network.

The integration of these techniques culminates in the development of sophisticated **recommender systems** that blend **Association Rule Learning (ARL)** and **Collaborative Filtering** for personalized product suggestions. This hybrid system leverages transactional data and user interactions to anticipate customer preferences, delivering tailored recommendations that enhance the shopping experience. By combining descriptive analytics with predictive capabilities, the project offers a robust framework for driving data-driven decision-making. Businesses can use these insights to optimize inventory management, improve customer retention, and maximize revenue through precision-targeted marketing strategies. Overall, this analysis not only provides actionable recommendations but also establishes a scalable model for leveraging customer data to achieve business growth and customer satisfaction.

Moreover, the project highlights the value of combining diverse analytical approaches to gain a comprehensive understanding of customer behavior. Each technique contributes uniquely to the analysis—association rule mining reveals purchasing interdependencies, clustering identifies customer personas, and dimensionality reduction refines the focus on impactful patterns. Social network analysis adds a relational perspective, uncovering connections between products that may not be immediately apparent through conventional methods. Together, these techniques create a holistic framework for understanding customer dynamics, enabling businesses to devise strategies that are both data-driven and customer-centric.

Beyond its technical and business implications, this project serves as a blueprint for leveraging analytics to bridge the gap between data and actionable insights. By integrating historical and real-time data, the model adapts to changing customer preferences, fostering continuous improvement in marketing, inventory management, and product development. This adaptive capability ensures that businesses are not only reacting to market demands but also anticipating them, positioning themselves as leaders in customer experience and operational efficiency. Ultimately, this multifaceted approach emphasizes the transformative potential of analytics in driving business success in an increasingly data-centric world.

8. USAGE

This framework is highly versatile and adaptable, making it applicable across a wide range of industries, including **e-commerce**, **retail**, **hospitality**, and more. By leveraging advanced analytics techniques, businesses can uncover actionable insights into customer behavior, enabling them to tailor their strategies for maximum impact. In **e-commerce**, for example, the framework can be used to analyze purchasing patterns, create personalized product recommendations, and implement dynamic pricing strategies that cater to individual customer preferences. Similarly, in **retail**, it can help businesses optimize store layouts, identify top-performing products, and design data-driven promotional campaigns to drive in-store and online sales. In the **hospitality industry**, this framework can be utilized to understand guest preferences, recommend services, and create personalized offers that enhance customer satisfaction and loyalty.

The insights derived from this framework empower businesses to design **targeted marketing campaigns** that resonate with specific customer segments, improving engagement and retention. By identifying influential products or services, businesses can focus their cross-selling and bundling strategies on items that drive the most value, boosting overall sales performance. Moreover, the ability to segment customers based on behavior allows for precision targeting, such as offering exclusive discounts to high-value customers or re-engagement offers to dormant ones. This level of personalization not only enhances the customer experience but also fosters brand loyalty, ensuring long-term growth and competitive advantage. Whether optimizing product recommendations, managing inventory, or crafting tailored promotions, this framework serves as a powerful tool for data-driven decision-making across industries.