Customer Transaction Pattern Analysis Using Data Mining: An Integrated K-Means and FP-Growth Approach

Daniel Osaroboh Atalor[1], Richard Odiase Okosodo[2]

[1]School of Computing, Engineering and Technology, Robert Gordon University, United Kingdom

[2]Department of Applied Artificial Intelligence, University of Greater Manchester, United Kingdom

DOI: https://doi.org/10.47772/IJRISS.2026.100500687

Received: 20 May 2026; Accepted: 25 May 2026; Published: 11 June 2026

ABSTRACT

This study addresses the fundamental challenge of leveraging high-volume transaction records to achieve simultaneous business goals: enhancing personalisation and driving strategic financial valuation. By applying a structured data mining methodology to the Online Retail II Dataset, this research successfully transformed complex transaction records into actionable strategic intelligence. The segmentation process involved IQR-based outlier exclusion and Standard Scaling of RFM features to mitigate data skew. This prepared data was then analysed using K-Means clustering for market segmentation and the FP-Growth algorithm for generating association rules. Key findings confirmed the Pareto Principle, showing that the top 27.6% of customers generate most of the revenue, justifying resource allocation toward high-value segments (VIP/Loyal Champions). Furthermore, the FP-Growth model generated exceptionally strong association rules (Lift up to 52.79), which were estimated to drive over $124,000 in incremental annual revenue through optimised cross-selling. The results validate the quantitative efficacy of integrating these data mining techniques to provide a financial quantifiable roadmap for optimising marketing and operational strategies in the modern digital economy.

Keywords: Data Mining, Customer Transaction Analysis, K-Means Clustering, FP-Growth Algorithm, RFM Analysis, Market Basket Analysis, Retail Analytics, Cross-Selling, Business Intelligence, Machine Learning, E-Commerce Analytics, Predictive Analytics.

INTRODUCTION

In the modern digital economy, companies have an unprecedented flow of customer transaction information from e-commerce portals, point-of-sale systems, and internet banking sites [1][2]. These large datasets, which are novel in their size and complexity, have the potential to unveil latent customer behaviours and anomalies. Nonetheless, manual examination is impractical, underscoring the need for sophisticated data mining methods to unlock actionable knowledge [3]. Examination of transaction patterns that can be made programmable through the application of algorithms such as clustering and association rule mining allows business firms to turn raw logs into strategic intelligence, thereby informing decisions in domains such as marketing personalisation [1] [4].

The relevance of such analysis depends on several conditions. First, the nature of transactions nowadays necessitates automated discovery to identify associations, such as popular purchase combinations or anomalies that would otherwise go undetected [3][4]. Second, against the backdrop of highly competitive markets, understanding customers' behaviour is integral to personalisation and customer retention; e.g., association-rule-powered recommendation systems have greatly enhanced the effectiveness of cross-selling [5] [6]. Finally, the development of regulatory frameworks, such as the GDPR, Nigeria's Data Protection Act of 2023, and the EU AI Act of 2025, necessitates that mining methods be both efficient and compliant.

Although data mining has been widely applied to customer transaction data, much of the existing literature remains limited in scope, often focusing on single-use cases such as customer segmentation [7] without demonstrating how transaction insights can address multiple business needs simultaneously, such as

personalisation, risk management, and inventory optimisation [1]. Research also tends to apply individual algorithms in isolation, such as Apriori for association rule mining [7] or K-means for clustering [8], with less emphasis on comparative or hybrid approaches that could enhance accuracy and interpretability in high-volume environments [4] [9].

On the shortcomings of prior work, most rely on batch processing of static datasets, overlooking the growing demand for scalable, real-time analytics in today's fast-moving digital economy [10] [11]. At the same time, emerging regulations, such as GDPR, highlight the need to embed compliance and ethical considerations into data-driven practices—an area still underexplored in technical mining studies [7] [13]. Finally, post-COVID shifts in consumer behaviour, including the adoption of mobile wallets, contactless payments, and Buy Now, Pay Later (BNPL) systems, have reshaped transaction dynamics in ways not yet fully captured by existing research [14] [15].

This paper addresses these gaps by analysing customer transaction patterns with selected data mining techniques, focusing on applications in personalisation, and discussing the implications of the resulting insights for scalability, real-time processing, and regulatory compliance in the 2025 data-driven economy. Building on this context, the study sets out to analyse customer transaction patterns using selected data mining techniques, with the aim of uncovering insights for personalisation and translating these findings into actionable business implications. To achieve this, the paper pursues two key objectives, which are further articulated through corresponding research questions.

1.1 Aim, Objectives, and Research Questions

Aim

This paper aims to analyse customer transaction patterns using data mining techniques to uncover insights for personalisation while considering implications for scalability, real-time processing, and compliance in the digital economy.

Objective

1. To apply selected data mining algorithms (e.g., clustering and association rule mining) to customer transaction data in order to uncover patterns relevant to personalisation
2. To interpret the discovered patterns to generate actionable business insights and discuss their implications for scalability, real-time analytics, and regulatory compliance.

Research Questions

- **RQ1:** How effective are selected data mining techniques in uncovering transaction patterns relevant to personalisation?
- **RQ2:** What actionable insights can be derived from these patterns, and what are their implications for scalability, real-time analytics, and regulatory compliance in the digital economy?

LITERATURE REVIEW

2.1 Customer's Transaction

Customer transactions are interactions between consumers and businesses, typically captured as digital records of purchases, payments, or exchanges [1]. These encompass a wide array of data points, including timestamps, item identifiers, purchase amounts, and customer demographics, enabling detailed behavioural profiling [2] [7].

2.1.1 Transaction Records

Transaction records are structured logs that include details such as timestamps, item identifiers, quantities, prices, payment methods, and customer demographics [1][2]. In modern systems, these records often integrate with location data, device information, and session metadata, forming comprehensive datasets for analysis [16] [17].

For instance, in retail, a record might detail a customer's purchase of multiple items in a single session, while in banking, it could track fund transfers or credit card usage over time [4]. The proliferation of digital platforms has generated billions of such records daily, necessitating advanced mining techniques to handle their volume and variety [3] [4].

2.1.2 Uses/Importance of Customer Transaction Records

Customer transaction records are pivotal for business success, offering multifaceted benefits. Primarily, they enable personalised marketing by revealing spending patterns, allowing companies to tailor promotions and improve customer satisfaction [7] [18]. In fraud detection, monitoring anomalies in transaction patterns helps prevent losses, with real-time analysis identifying suspicious activities like unusual purchase frequencies [19] [12]. Records also support inventory optimisation by forecasting demand based on historical purchases, reducing overstock and stockouts [16][20]. Furthermore, they facilitate customer segmentation for targeted strategies, enhancing retention and profitability [8][21]. Accurate record-keeping ensures financial transparency, supports regulatory compliance, and provides audit trails to deter identity theft [22][23]. Overall, these records inform strategic decisions, driving revenue growth and operational efficiency in competitive markets [2].

2.2 Data Mining

Data mining is the process of discovering patterns, correlations, and anomalies within large datasets using statistical, machine learning, and database techniques [1]. It is particularly adept at handling transactional data and evolving through integrations, such as AI, to enhance predictive capabilities [24].

2.2.1 Data Mining Algorithms

Several algorithms are tailored for transaction data mining, focusing on pattern discovery and prediction. Association rule mining identifies frequent item sets in transactions, such as products commonly bought together. This field was traditionally exemplified by the Apriori algorithm, but recent enhancements that address computational efficiency for large datasets have led to algorithms such as FP-Growth. FP-Growth eliminates the resource-intensive step of candidate generation, making it the preferred method for mining transactional databases [5].

Clustering techniques, like K-Means and hierarchical clustering, group similar transactions for segmentation [25]. Supervised methods, including decision trees and support vector machines (SVMs), are widely used to classify behaviours for churn prediction [26]. Recent developments include hybrid models, such as the Light Gradient Boosting Machine (LGBM) combined with XGBoost for fraud detection, which achieve high accuracy on imbalanced transaction datasets [27]. Neural networks and random forests handle complex, high-dimensional data [24]. Finally, challenges like pattern redundancy are addressed through targeted and heuristic mining, emphasising utility over mere frequency [30]

2.2.2 Data Mining Applications

Data mining plays a pivotal role in analysing customer transaction patterns, offering diverse applications that enhance business decision-making in the data-driven landscape of 2025. In e-commerce, recommendation systems, such as Amazon’s collaborative filtering, leverage transaction histories to suggest products, boosting sales by up to 35% in some cases [31]. Customer segmentation employs supervised techniques such as support vector machines (SVMs) and random forests, as well as unsupervised methods such as K-means, to group users by demographics, preferences, or behaviour, enabling targeted marketing strategies [12]. Fraud detection uses anomaly detection to identify irregular transaction patterns, mitigating financial risks [19]. Churn prediction, critical for customer retention, applies classification algorithms to forecast attrition and achieves high accuracy in identifying at-risk customers [21] [32][33].

Market basket analysis, driven by association rule mining, optimises promotions by identifying frequently co-purchased items, enhancing cross-selling opportunities [5] [34]. Social media-integrated data mining extracts insights for competitor analysis and loyalty programs, leveraging network analysis and natural language

processing (NLP) to classify sentiment and detect fake reviews [35][36][37]. In addition, e-commerce applications include user preference modelling, discrete choice analysis, and discount promotion misuse detection, utilising techniques like hidden Markov models, Bayesian models, and clustering [38][39].

Emerging trends focus on real-time applications, such as dynamic pricing and predictive maintenance in retail, supported by streaming data analysis and computer vision for enhanced customer profiling [10] [40]. These applications underscore the versatility of data mining in addressing complex, dynamic transaction datasets, driving personalisation, operational efficiency, and risk management in modern business environments.

2.3 Research Gaps and Study Rationale

Despite the growing use of data mining in the retail sector, much of the existing research remains fragmented and limited in scope. Many previous studies focus on single-use applications, such as isolated customer segmentation or fraud detection, without demonstrating how multiple business objectives, such as personalisation, cross-selling, and risk identification, can be achieved within a unified analytical framework. This study bridges that gap by integrating both K-Means and FP-Growth algorithms to generate holistic insights that serve multiple strategic functions simultaneously.

Another major gap in the literature lies in methodological efficiency and scalability. Prior research often applies traditional algorithms such as Apriori in isolation, neglecting more efficient alternatives like FP-Growth, which are better suited to the large, sparse datasets typical of modern retail. This study addresses that limitation by adopting computationally efficient techniques that enhance model performance and ensure practical applicability across high-volume environments.

Finally, while much of the existing work relies on static historical data, the findings of this study support real-time analytics, which are essential for modern digital operations. Furthermore, unlike many technical studies that focus solely on algorithmic output, this research aligns data-driven insights with strategic and compliance considerations. By embedding scalability and GDPR-aligned data practices, it ensures that analytical solutions are not only effective but also responsible and adaptable to the evolving regulatory and operational landscape

This study utilises an integrated K-Means and FP-Growth approach on real-world data to address these gaps, demonstrating how a single methodology delivers high-value intelligence for multiple simultaneous business objectives.

METHODOLOGY

The proposed methodology employs a structured, iterative data-mining framework to analyse customer transaction patterns, delivering insights for personalisation, fraud detection, and strategic recommendations in the dynamic business environment of 2025 [1]. It is designed to handle large-scale, complex transaction datasets and address scalability and real-time processing needs. The Methodology section covers data collection, cleaning, EDA, feature engineering, Outler detection and handling and model application.

3.1 Data Collection

The study utilises the Online Retail II Dataset, sourced from a real UK-based retailer specialising in unique giftware, covering approximately two years. The dataset is substantial, encompassing 1,067,371 individual transactions conducted across 38 different countries. The scale of the retail operation is evidenced by the involvement of 5,942 unique customers who purchased from a catalogue of 4,223 unique products, providing a rich basis for analysing customer behaviour, purchase patterns, and cross-national trends. Although the Online Retail II Dataset provides a large and rich dataset for customer transaction analysis, using a single retail dataset limits the generalisability of the findings to other sectors, such as banking, telecommunications, grocery retail, and hospitality. Therefore, the results should be interpreted within the context of giftware retail transactions. Future validation using multiple datasets from different industries would strengthen the robustness and external validity of the proposed framework.

Table 3.1 presents the attribute descriptions for the dataset. These eight attributes are crucial for analysing transaction patterns, customer behaviour, and potential anomalies, and they form the foundation of the study's data-mining objectives.

Table 3.1 Attributes Description of the Dataset

Attribute	Data Type	Description and Purpose	Relevance to Study
Invoice No	Nominal	Unique 6-digit transaction identifier. A leading 'c' denotes a cancellation, which is critical for fraud/anomaly analysis.	Transaction Grouping & Fraud Detection
Stock Code	Nominal	Unique 5-digit product identifier.	Item-level Pattern Analysis
Description	Nominal	The name of the product item.	Personalisation & Item Description
Quantity	Numeric	The number of units purchased per item in the transaction.	Transaction Value & Outlier Detection
Invoice Date	Numeric (Timestamp)	Date and time the transaction was generated.	Temporal Analysis & Real-Time Analytics
Unit Price	Numeric	The price per product unit in sterling (£)	Transaction Value Calculation
Customer ID	Nominal	Unique 5-digit customer identifier.	Customer Segmentation & Personalisation
Country	Nominal	The customer's country of residence.	Geographic Segmentation & Compliance Discussion

The dataset tracks individual retail line items. Transactions are identified by Invoice number and products by Stock Code and description. Financial details include quantity and unit Price. Every purchase is time-stamped with the Invoice Date, and customers are uniquely tracked by Customer ID and Country. Initial records confirm that single customers often purchase multiple distinct items simultaneously, revealing typical market basket contents.

3.2 Data Cleansing

For data cleansing and anomaly handling, missing-value imputation, transaction validity checks, and cancellation management were performed to ensure data integrity and analytical accuracy.

The data was cleaned to remove invalid or missing entries. Additional preprocessing operations, including feature creation (e.g., Revenue calculation) and date extraction, were also performed using pandas' built-in methods. This ensured systematic, reproducible, and efficient data cleaning and transformation. Data cleaning involved removing records without a Customer ID, ensuring valid inputs for clustering; enforcing transaction validity by excluding entries with non-positive Quantity or Unit Price after computing Total Sales = Quantity × Unit Price; and isolating cancelled transactions (Invoice No starting with "C") for separate anomaly analysis, while excluding them from segmentation and ARM to retain only standard purchase patterns

3.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted on the cleaned dataset to generate statistical insights guiding subsequent modelling. Six visualisations were used to examine key dimensions: Temporal Patterns, revealing strong year-end seasonality and mid-week sales peaks; Geographic Concentration, identifying the United Kingdom as the primary revenue source and indicating the need for country-specific treatment; and Feature Skewness, where highly skewed Quantity and Price distributions justified aggressive outlier handling prior to K-Means clustering.

The visualisations for EDA were generated by grouping Total Sales by temporal variables (Hour, Month, Day of Week) and the categorical variable (Country), and then calculating frequency distributions (histograms) for Price and Quantity.

3.3.1 Descriptive Analysis and KPI

Descriptive statistical analysis was carried out using basic aggregation and counting techniques applied to the cleaned dataset. Total metrics were derived by counting unique entries for customers, products, and countries, and summing *Total Sales* to obtain the overall Total Revenue. Additionally, the Average Order Value (AOV) was computed as the ratio of Total Revenue to the total number of unique invoices, providing insight into the average monetary value of individual customer transactions.

3.3.2 Outlier Detection and Handling

Outliers were detected and removed to enhance the accuracy of the K-Means clustering. Using the Interquartile Range (IQR) method, records with Monetary Value or Frequency beyond (Q1 − 1.5 × IQR) or (Q3 + 1.5 × IQR) were excluded. This ensured the resulting segments reflected genuine customer behaviour, free from distortion by extreme or wholesale-driven values.

Table 3.2 Descriptive Statistics for RFM Features

Statistic	Customer ID	Monetary Value (£)	Frequency (Orders)	Last Invoice Date	Recency (Days)	Count
Count	5,167	5,167	5,167	5,167	5,167	5,167
Mean	15,341.72	1,133.52	3.69	2011-05-04 02:42:25	218.90	5,167
Standard Deviation	1,705.11	1,128.97	3.19	–	211.12	–
Minimum	12,348.00	2.95	1.00	2009-12-01 10:49:00	0.00	5,167
25% (Q1)	13,862.50	311.03	1.00	2010-11-11 13:06:30	32.00	5,167
50% (Median)	15,355.00	717.51	3.00	2011-07-28 14:08:00	133.00	5,167
75% (Q3)	16,811.50	1,579.13	5.00	2011-11-06 13:22:30	392.00	5,167
Maximum	18,287.00	5,202.97	16.00	2011-12-09 12:50:00	738.00	5,167

The descriptive statistics indicate strong positive skewness in Monetary Value (£1,133.52 mean vs £717.51 median) and Frequency (3.69 mean vs 3.00 median), suggesting the presence of high-spending outliers, likely wholesale or bulk buyers. The wide range of £2.95 to £5,202.97 in Monetary Value and 1 to 16 orders in Frequency further confirms this. Additionally, the high standard deviations (£1,128.97 for Monetary Value and 3.19 for Frequency) reinforce the variability within the dataset. These findings validate the need for IQR-based outlier removal to ensure the K-Means clustering reflects genuine consumer behaviour rather than extreme anomalies

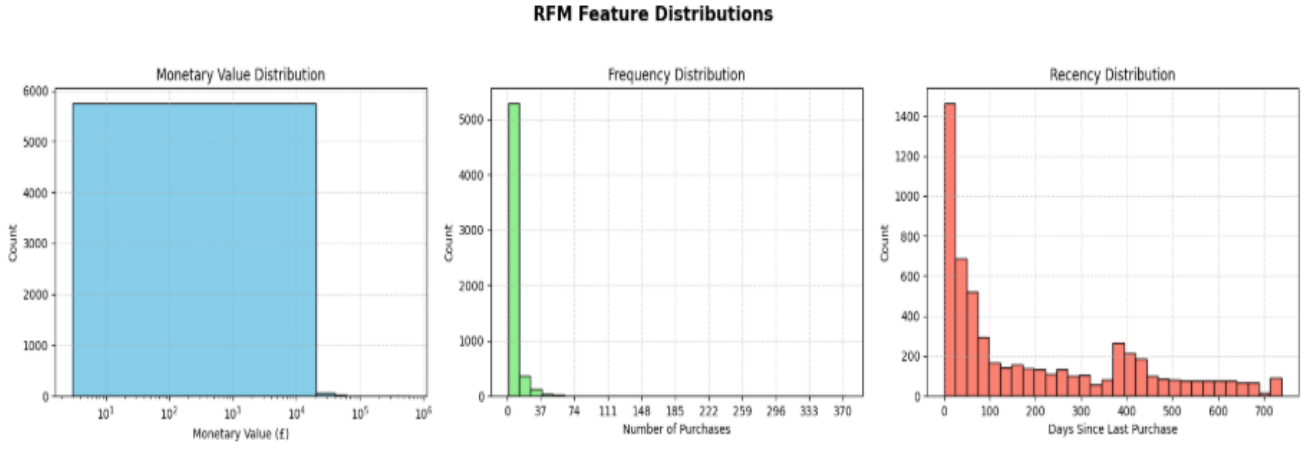

Figure 3. 1 RFM Feature Distribution

The analysis confirms extreme positive skewness in both Monetary Value and Frequency, thereby validating the need for aggressive outlier exclusion using the IQR method. This skew is caused by a few ultra-high-value

customers, and removing them ensures the clustering focuses on the core customer base. The Recency profile supports this, showing a large group of recent buyers and a long tail of dormant customers, which requires creating distinct segments for immediate action versus win-back campaigns.

3.4. Feature Engineering

To facilitate customer segmentation, the transactional data was aggregated and synthesised into three key behavioural metrics using the Recency, Frequency, Monetary (RFM) framework.

The RFM (Recency, Frequency, and Monetary Value) model was applied to evaluate customer purchasing behaviour and to segment customers. Recency (R) was calculated as the number of days between the latest recorded date in the dataset (the cut-off date) and each customer's most recent Invoice Date, indicating how recently a customer made a purchase. Frequency (F) represented the total count of unique Invoice No linked to each Customer ID, reflecting the customer's transactional volume. Lastly, Monetary Value (M) was calculated as the cumulative sum of Total Sales for each Customer ID, providing insight into the customer's overall financial contribution.

Table 3. 3 Customer Feature Matrix (RFM Metrics)

Customer ID	Monetary Value (£)	Frequency (Orders)	Last Invoice Date	Recency (Days)
12,346.00	77,352.96	3	2011-01-18 10:01:00	325
12,347.00	5,633.32	8	2011-12-07 15:52:00	2
12,348.00	1,658.40	5	2011-09-25 13:13:00	74
12,349.00	3,678.69	3	2011-11-21 09:51:00	18
12,350.00	294.40	1	2011-02-02 16:01:00	309

3.4.1. Data Transformation (ARM)

The data was structured for market basket analysis by converting the line-item format into a sparse transaction matrix.

One-Hot Encoding was applied to transform the dataset into a machine-learning-compatible format for association rule mining (ARM). The data was pivoted using *Invoice No* as the transaction identifier and *Stock Code* as the item identifier. The resulting matrix was then one-hot encoded, with a value of '1' indicating the presence of a product on a specific invoice and '0' indicating its absence. This sparse, binary-encoded structure provided the necessary input format for the chosen ARM algorithm, such as Apriori, enabling the discovery of meaningful associations among products.

3.4.2. Final Data Scaling

Prior to applying the clustering algorithms, the final RFM feature set underwent Standardization (Z-score scaling). This was a crucial step that ensured all features were centred on a mean of zero with a standard deviation of 1. This scaling prevents features with naturally larger magnitudes (such as Monetary Value) from disproportionately influencing the partition-based clustering process, ensuring that the K-Means algorithm measures distances fairly across the Recency, Frequency, and Monetary dimensions. Following this, the pre-processed data was divided into two distinct analytical subsets: a standardised RFM feature matrix, which served as the final input for K-Means clustering, and a sparse, one-hot-encoded transaction matrix, which was utilised for FP-Growth association rule mining.

3.5 Model Application

3.5.1 Association Rule Mining (FP-Growth)

The methodology used in generating association rules was the FP-Growth algorithm. It's chosen for its efficiency in market basket analysis as it eliminates the slow candidate-generation step required by Apriori.

The process was governed by two key thresholds: a minimum support of 1.0% and a minimum confidence of 30.0%. The algorithm first constructed an FP-Tree to efficiently identify all frequent item sets within the transaction data. These item sets were then transformed into "If–Then" rules (X → Y), which were retained only if their confidence exceeded the specified threshold. Later, the lift metric was applied to evaluate the strength and practical relevance of each rule, ensuring that only statistically meaningful and business-relevant associations were included in the final output. Despite its efficiency, FP-Growth also has limitations. The quality of the rules depends on the selected support and confidence thresholds; very low thresholds generate too many rules, including weak patterns, while high thresholds exclude useful but less frequent associations. FP-Growth also identifies statistical co-occurrence rather than true customer motivation. Therefore, the discovered rules should be interpreted as evidence of product association, not proof that one product causes the purchase of another. Future work should compare FP-Growth with Apriori, ECLAT, and utility-based association rule mining to assess rule quality, scalability, and commercial usefulness.

3.5.2 K-Means

K-Means was used for customer segmentation primarily because it is an efficient, interpretable, and widely accepted algorithm for partition-based clustering, and is especially well-suited to the Recency, Frequency, Monetary (RFM) feature set. Figure 3. 2 displays two common metrics used to determine the optimal number of clusters for the K-Means algorithm: the Elbow Method (Inertia) and the Silhouette Score. The visual evidence justifies using k=4.

However, K-Means has important assumptions and limitations. It assumes that clusters are relatively spherical, separable, and similar in size, which may not always reflect real customer behaviour. The algorithm is also sensitive to the chosen number of clusters, the initialisation of centroids, outliers, and feature scaling. To address these weaknesses, this study applied IQR-based outlier handling, standard scaling, and both the Elbow Method and the Silhouette Score to justify selecting four clusters. Nevertheless, future studies should compare K-Means with alternative clustering methods such as Hierarchical Clustering, DBSCAN, Gaussian Mixture Models, and Self-Organising Maps.

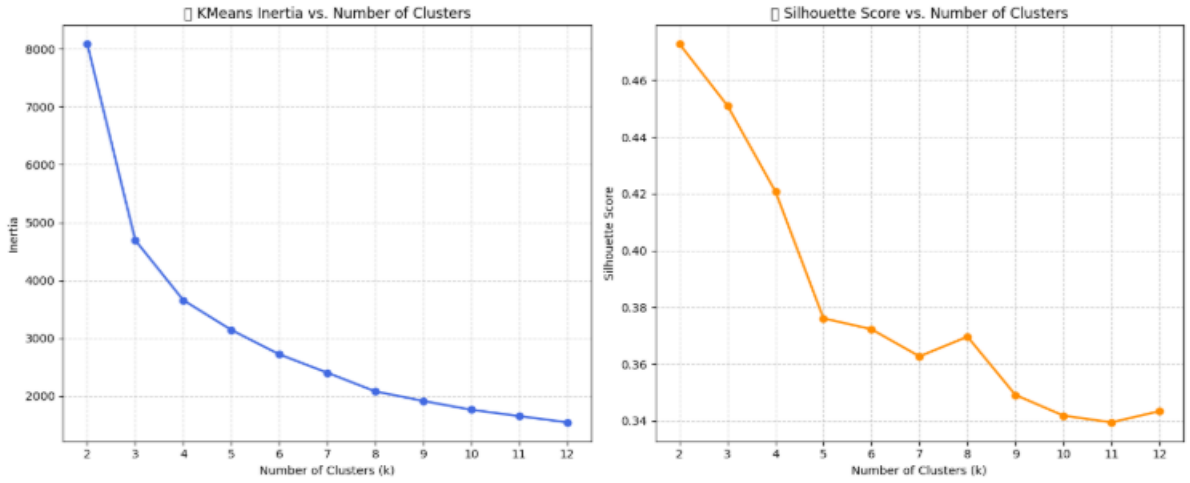

Figure 3. 2 Determine the number of clusters

The K-Means clustering process was executed in a structured, multi-step sequence to ensure accurate customer segmentation. First, the RFM features—Total Revenue, Number of Orders, and Days Since Last Order—required Standard Scaling (Z-score) to neutralise scale differences, a critical prerequisite, as the algorithm relies on Euclidean distance and is sensitive to feature magnitudes. The K-Means algorithm then iteratively clustered the scaled data into four predetermined groups k =4 by randomly initialising centroids, assigning each customer to the nearest centroid, and updating the centroid's position to the mean of its members until the cluster assignments stabilised. Following convergence, the final segmentation was analysed by averaging the original (unscaled) RFM metrics for each of the four resulting groups, thereby defining the characteristic profile (e.g., Average Revenue, Average Orders) of each customer segment for direct use in personalised marketing strategies.

3.5.3 Implementation and Scalability

The methodology was implemented using Python, leveraging foundational libraries essential for data mining and processing. Specifically, pandas facilitated data cleaning and feature engineering (RFM metric creation), while scikit-learn provided the framework for the K-Means clustering and Standard Scaling. The FP-Growth algorithm was executed using specialised association rule libraries. This approach provides a direct path to generating actionable insights (e.g., cross-selling strategies, customer segmentation).

Table 3.4 Justification of Selected Algorithms

Analytical Task	Selected Method	Alternative Methods	Justification
Customer segmentation	K-Means	Hierarchical Clustering, DBSCAN, Gaussian Mixture Models	K-Means was selected because it is efficient, interpretable, and suitable for standardised RFM features.
Association rule mining	FP-Growth	Apriori, ECLAT, Utility-Based ARM	FP-Growth was selected because it avoids candidate generation and is more efficient for large, sparse transaction datasets.
Customer value analysis	RFM	CLV modelling, behavioural scoring	RFM was selected because it provides simple, interpretable measures of recency, frequency, and monetary contribution.

This comparison shows that the chosen methods were suitable for the study objectives, while other alternative techniques offer further insights for future research.

RESULTS AND DISCUSSIONS

This section presents the empirical findings of the data mining process. It begins with dataset validation and an initial descriptive analysis, followed by the presentation and interpretation of the two core models: K-Means clustering for customer segmentation and FP-Growth for association rule mining. The subsequent discussion synthesises these findings, compares them to relevant literature, and quantifies the estimated financial value derived from the actionable business intelligence.

4.1 Dataset

This analysis begins by validating the dataset's characteristics and integrity, summarising the attributes used and detailing the data quality assessment that guided the subsequent preprocessing. A sample of the data is shown in Table 4.1

Table 4. 1 Sample of Raw Online Retail Transaction Data

Index	Invoice	Stock Code	Description	Quantity	Invoice Date	Price	Customer ID	Country
0	489434	85048	15cm Christmas Glass Ball 20 Lights	12	2009-12-01 07:45:00	6.95	13085.00	United Kingdom
1	489434	79323P	Pink Cherry Lights	12	2009-12-01 07:45:00	6.75	13085.00	United Kingdom
2	489434	79323W	White Cherry Lights	12	2009-12-01 07:45:00	6.75	13085.00	United Kingdom
3	489434	22041	Record Frame 7" Single Size	48	2009-12-01 07:45:00	2.10	13085.00	United Kingdom
4	489434	21232	Strawberry Ceramic Trinket Box	24	2009-12-01 07:45:00	1.25	13085.00	United Kingdom

5	489434	22064	Pink Doughnut Trinket Pot	24	2009-12-01 07:45:00	1.65	13085.00	United Kingdom
6	489434	21871	Save The Planet Mug	24	2009-12-01 07:45:00	1.25	13085.00	United Kingdom
7	489434	21523	Fancy Font Home Sweet Home Doormat	10	2009-12-01 07:45:00	5.95	13085.00	United Kingdom
8	489435	22350	Cat Bowl	12	2009-12-01 07:46:00	2.55	13085.00	United Kingdom
9	489435	22349	Dog Bowl, Chasing Ball Design	12	2009-12-01 07:46:00	3.75	13085.00	United Kingdom

The dataset tracks individual retail line items, with transactions identified by the invoice number and products by their stock code and description. The volume of items purchased is captured by quantity, and the price per unit is recorded in price. Each transaction is time-stamped with the Invoice Number, and products are identified by their Stock Code and Description. Customers are uniquely tracked by Customer ID, and their geographic location is noted by Country.

A comprehensive data quality assessment was carried out to evaluate the completeness, accuracy, and consistency of the dataset. Table 4.2 shows the information on data quality.

Table 4. 2 Data Quality Report

Column	Type	Non-Null	Null	Null %	Unique	Sample	Quality
Invoice	Object	1,067,371	0	0.0%	53,628	489434	Good
Stock Code	Object	1,067,371	0	0.0%	5,305	85048	Good
Description	Object	1,062,989	4,382	0.4%	5,698	15cm Christmas glass ball 20 lights	Good
Quantity	int64	1,067,371	0	0.0%	1,057	12	Good
InvoiceDate	Object	1,067,371	0	0.0%	47,635	2009-12-01 07:45:00	Good
Price	float64	1,067,371	0	0.0%	2,807	6.95	Good
Customer ID	float64	824,364	243,007	22.8%	5,942	13085.0	Warning
Country	Object	1,067,371	0	0.0%	43	United Kingdom	Good

The results indicated that most variables demonstrated strong data integrity, with over 99% completeness across key fields. However, notable issues were identified, including 22.8% missing Customer IDs, 22,950 negative quantities (likely representing returns), 6,202 zero-priced items, and 19,494 cancelled orders. Tables 4.3 and 4.4 show the information on data cleaning and summary.

Table 4.3 Data Cleaning Summary

Category	Action/Description	Count
Cancelled Orders	Removed	19,494
Returns	Removed	3,457
Invalid Prices	Removed	2,750
Revenue Column	Created	—
Date Components	Extracted	—

Table 4.4 Data Reduction Overview

Stage	Row Count	Change	Percentage Removed
Original	1,067,371	—	—
Cleaned	1,041,670	-25,701	2.4%

Data cleaning and reduction were conducted to enhance dataset reliability and analytical accuracy. The process involved removing 19,494 cancelled orders, 3,457 returns, and 2,750 invalid price entries, and creating new Revenue and date component fields to support further analysis. After cleaning, the dataset was reduced from 1,067,371 to 1,041,670 records, representing a 2.4% reduction and ensuring that only valid and meaningful transactions were retained.

4.2 Exploratory Data Analysis (EDA)

Figure 4.1 presents the results of the Exploratory Data Analysis (EDA), illustrating findings across three key analytical areas: Temporal Patterns, Geographic Concentration, and Feature Skewness. The Exploratory Data Analysis confirms the data's suitability for advanced modelling while revealing critical characteristics that necessitate specific preprocessing.

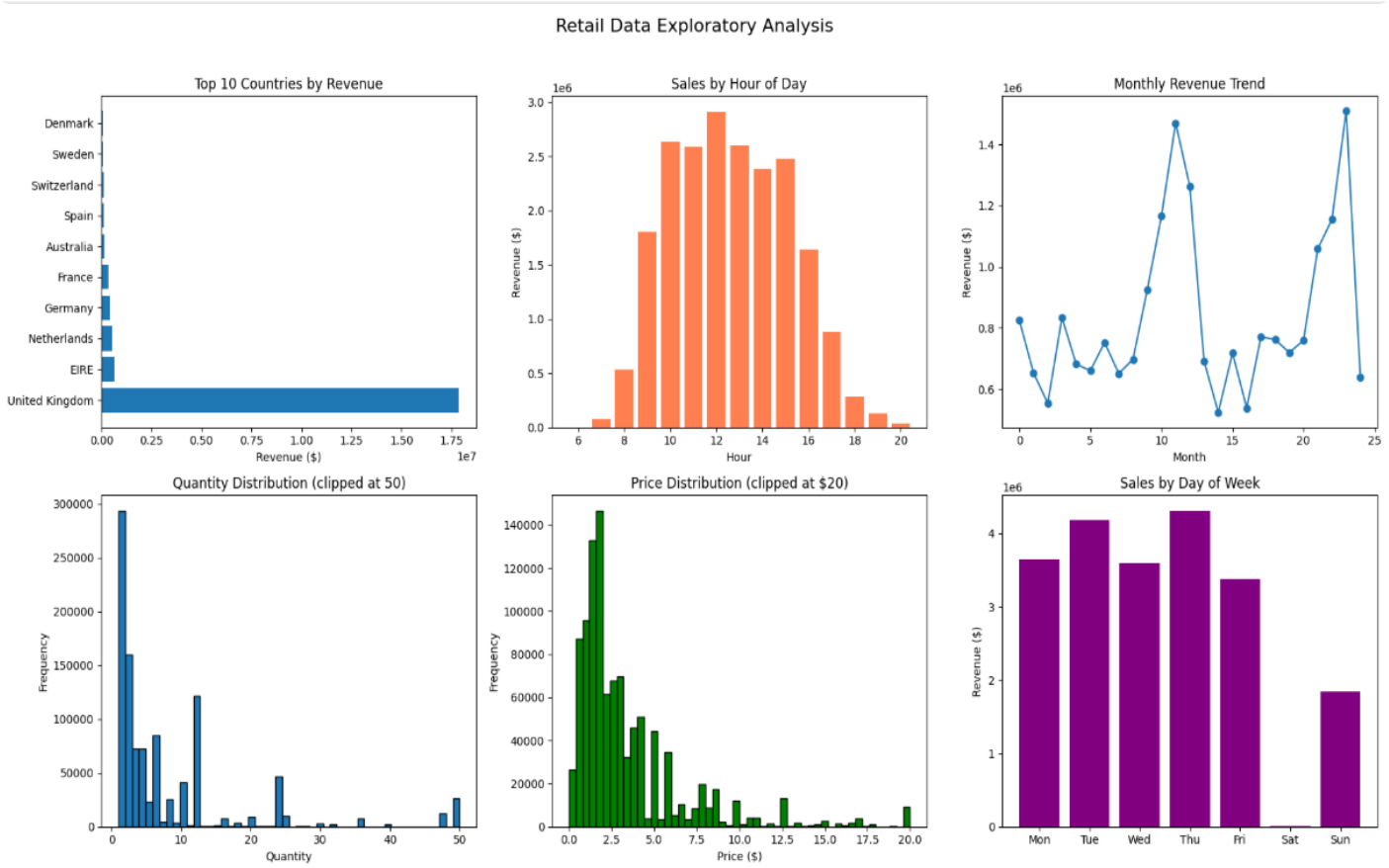

Figure 4.1 Retail Data Exploratory Analysis

The "Top 10 Countries by Revenue" chart confirms extreme revenue dependence, with the United Kingdom vastly dominating sales. This justifies the need for standardisation during preprocessing to avoid UK-scale bias in the distance metrics used by K-means.

The analysis of the Monthly Revenue Trend highlights strong year-end seasonality, while the Sales by Hour and Sales by Day of Week charts identify critical midday (11:00-14:00) and midweek sales peaks, providing context for real-time analytics and wholesaler behaviour.

Histograms of Quantity and Price show severe positive skewness, with a pronounced concentration near zero. This empirical observation confirms the need for logarithmic transformation of the derived aggregate features (Monetary Value and Frequency) to stabilise variance and ensure valid performance during K-means clustering.

4.3 Statistics

Initial classical descriptive statistics were used to characterise the population (customers, products) and to measure key performance indicators (KPIs) such as Revenue and AOV. Table 4.5 shows the descriptive statistics

of the dataset.

Table 4. 5 Descriptive Statistics

Metric	Value
Total Revenue	$20,972,594.57
Total Customers	5,878
Total Products	4,917
Total Countries	43
Average Order Value (AOV)	$20.13

It can be seen that the retailer is a financially substantial operation with $20.97 Million in Total Revenue, a high customer count of 5,878, and a broad catalogue of 4,917 products across 43 countries, confirming the dataset's suitability for large-scale data mining (FP-Growth). The relatively low Average Order Value of $20.1, despite the large customer base, reinforces the EDA finding of a high-volume, low-value transactional focus, which emphasises the importance of analysing the Frequency RFM metric and leveraging FP-Growth to optimise cross-selling strategies for increased basket size.

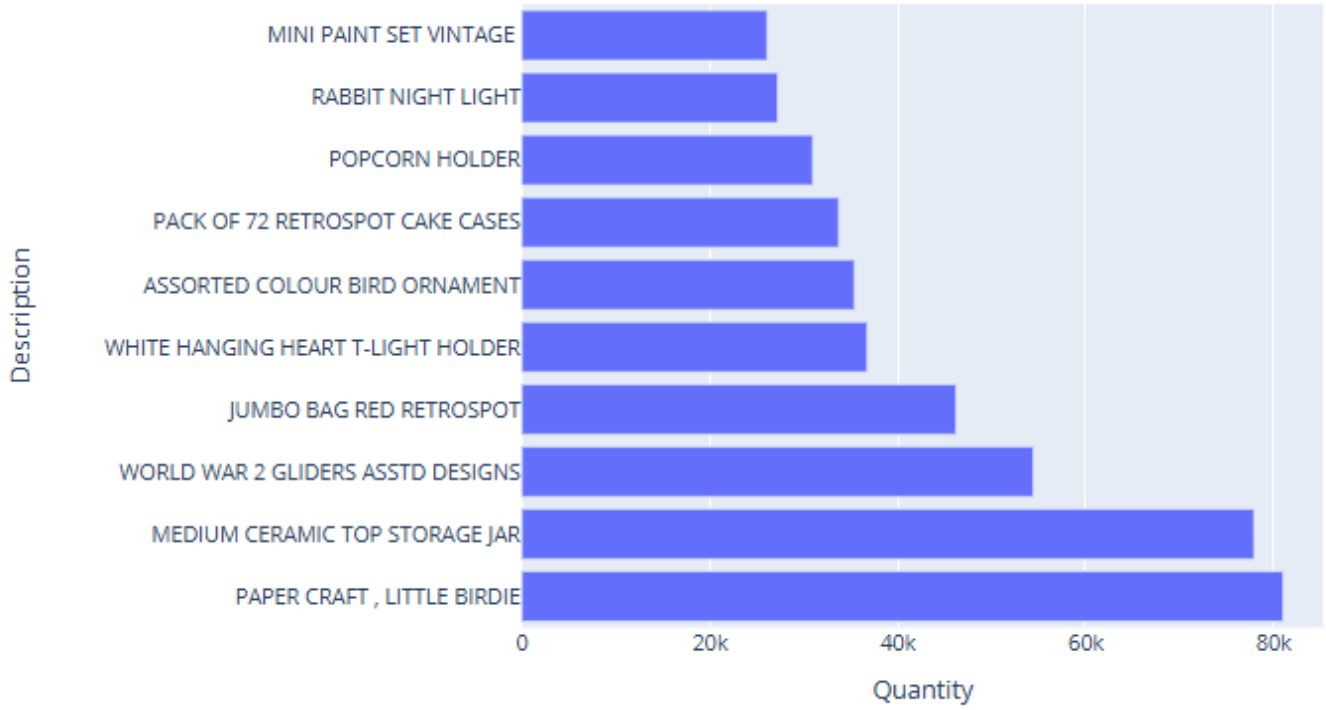

Figure 4.2 Top 10 Products by Quantity Sold

The top 10 products by quantity sold are dominated by high-volume, low-cost items such as paper craft, storage jars, and small giftware. The two clear leaders, 'paper craft, little birdie' and 'medium ceramic top storage jar' (each selling around 78k units), drive overall sales volume significantly. This high transactional velocity highlights the critical need for robust inventory management and stocking priority for these specific items to support the retailer's sales frequency.

4.4 Feature Engineering

Following the initial data cleanup (the resulting state of the data before modelling), the RFM Feature Distributions (Figures 4.8 and 4.9) show persistent skewness after outlier removal, confirming the need for Standard Scaling before clustering.

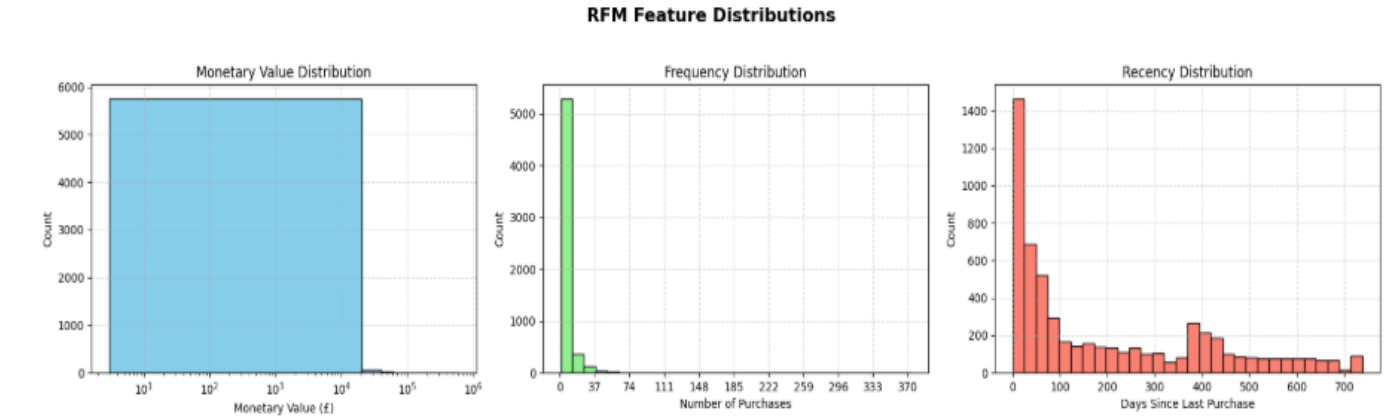

Figure 4.3 RFM Feature Distribution

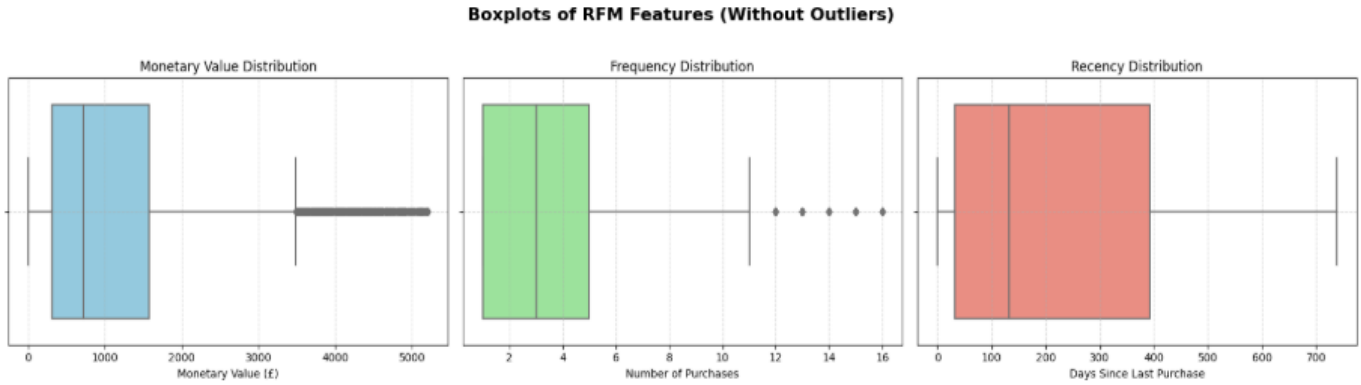

Figure 4. 4 Boxplot of RFM Features

The analysis confirms extreme positive skewness in both Monetary Value and Frequency, thereby validating the need for aggressive outlier exclusion using the IQR method. This skew is caused by a few ultra-high-value customers, and removing them ensures the clustering focuses on the core customer base. The Recency profile supports this, showing a large group of recent buyers and a long tail of dormant customers, which requires creating distinct segments for immediate action versus win-back campaigns.

The boxplots confirm that the IQR outlier filtering successfully removed the most extreme, non-representative values, but the data remains heavily skewed. This persistent positive skewness in both Monetary Value and Frequency justifies the subsequent, necessary preprocessing step: Standard Scaling (Z-score), which will ensure that these features contribute equally to the distance calculations in the K-Means model. Meanwhile, the wide distribution of Recency provides a clear basis for successfully segmenting customers into distinct groups, ranging from recent active buyers to dormant users.

4.5 Model Application

4.5.1 Association Rule Mining Outcomes

Table 4.7 presents the results of the FP-Growth algorithm, summarising the scale and distribution of frequent item sets identified across various lengths under the defined thresholds.

Table 4.6 Key Association Rules for Personalisation and Cross-Selling

Parameter/Metric	Value
Total Transactions Analysed	5,000
Minimum Support	1.0%

Minimum Confidence	30.0%
Frequent 1-Itemsets Found	836
Frequent 2-Itemsets Found	1,383
Frequent 3-Itemsets Found	335
Frequent 4-Itemsets Found	33
Frequent 5-Itemsets Found	1

The analysis was conducted on a dataset of 5,000 transactions using relatively low thresholds, specifically a minimum support of 1.0% and a minimum confidence of 30.0%. These settings were intentionally chosen for exploratory purposes to capture a wide range of potential co-occurrence patterns. The algorithm identified 2,538 frequent item sets (836 single-item sets, 1,383 two-item sets, 335 three-item sets, 33 four-item sets, and 1 five-item set), reflecting a rich diversity of products frequently purchased together.

The distribution pattern aligns with the expected sparsity of retail data. The largest concentration of patterns occurred among 2-itemsets (1,383), indicating that product pairs represent the most frequent and actionable co-purchase relationships. This was followed by a sharp decline in higher-order combinations; 335 at length 3, and only 1 at length 5, confirming that multi-item combinations beyond triplets are rare. The findings validate the effectiveness of the FP-Growth algorithm in detecting meaningful co-purchase relationships. The results suggest that cross-selling and recommendation strategies should focus primarily on 2- and 3-item combinations, as these represent the most statistically significant and commercially valuable patterns within the dataset.

Table 4. 7 Association Rules Summary Table

Rule ID	Antecedent (IF)	Consequent (THEN)	Support	Confidence (%)	Lift
2072	Envelope 50 Blossom Images	Pack of 72 Retro Spot Cake Cases, Envelope 50 Romantic Image	0.015	67.6	52.79
2076	Pack of 72 Retro Spot Cake Cases, Envelope 50 Romantic Image	Envelope 50 Blossom Images	0.015	78.1	52.79
2070	Dotcom Postage, Envelope 50 Romantic Images	Envelope 50 Blossom Images	0.014	77.5	52.34
825	Metal Sign Cupcake Single Hook	Metal Sign, Cupcake Single Hook	0.017	84.1	50.08
2502	Charlotte Bag, Pink/White Spots, Pack of 72...	Red Spotty Charlotte Bag, Dotcom Postage	0.016	75.8	47.95
2491	Dotcom Postage, Charlotte Bag, Pink/White Spots	Red Spotty Charlotte Bag, Pack of 72 Retro Spot Cake Cases	0.015	64.1	47.84

The key association rules found are highly specific and exceptionally strong, confirming genuine customer buying habits rather than random chance. The extremely high Lift values (up to 52.79) indicate that when a customer buys the "IF" item(s), they are over 50 times more likely to buy the "THEN" item(s) than would be expected by chance, demonstrating strong predictive utility. These rules often link tightly complementary products (such as specific envelopes, cake cases, and metal signs) with high Confidence (up to 84.1%), making them directly actionable for recommendation systems and inventory bundling.

4.5.2 Customer Behavioural and Contextual Interpretation

While the quantitative results reveal strong purchasing patterns, highlighting behavioural and contextual factors shows their importance in understanding customer decisions. For example, seasonal demand, gift-buying behaviour, product aesthetics, discounts, stock availability, and customer loyalty may explain why certain products are frequently purchased together. The strong association rules involving giftware-related items may reflect occasion-based purchasing, where customers buy complementary items for celebrations, home decoration, or gifting purposes. Similarly, dormant customers may not only represent dissatisfaction or churn

risk, but may also reflect seasonal buyers, one-off gift purchasers, or customers whose needs are irregular. Therefore, the results should be interpreted alongside the customer context rather than numerical patterns.

4.5.3 K means

The K-Means algorithm successfully segmented the customer base into four distinct groups, allowing for targeted personalisation strategies. Clusters were defined using the mean values of the RFM-derived features: total revenue (monetary), number of orders (frequency), and days since last order (recency, where lower values indicate higher engagement). Table 4. 8 shows the Characteristics of K-Means Customer Segments.

Table 4. 8 Characteristics of K-Means Customer Segments (RFM Profile)

Cluster ID	Size (Customers)	Avg Revenue (Monetary)	Avg Orders (Frequency)	Avg Days Since Last Order (Recency)	Business Label
0	1,005	$2,247.44	6.4	93.1	High-Value/Active
1	447	$3,826.49	13.0	56.7	VIP/Loyal Champions
2	2,071	$644.93	$2.5	$93.0	Low-Value/Risk of Churn
3	1,727	$509.33	1.8	482.4	Dormant/Lost

Interpretation

The RFM clustering analysis revealed four distinct customer segments, each exhibiting unique behavioural and monetary characteristics that inform tailored marketing strategies.

Cluster 1 (VIP/Loyal Champions), although the smallest group with 447 customers, represents the most valuable segment. These customers generate the highest average revenue ($3,826.49), exhibit the highest purchase frequency (13.0 orders), and have the lowest recency (56.7 days), making them ideal candidates for retention and reward initiatives.

Cluster 0 (High-Value/Active) comprises a large group of high-spending customers with an average revenue of $2,247.44, who should receive personalised offers to sustain engagement and loyalty.

Cluster 2 (New/Risk of Churn) is the largest segment, with 2,071 customers, but shows low monetary and frequency scores, indicating the need for immediate activation strategies to encourage repeat purchases.

Cluster 3 (Dormant/Lost) includes customers who have not made a purchase for an average of 482 days, representing the least valuable group that should be approached only with targeted, high-incentive win-back campaigns.

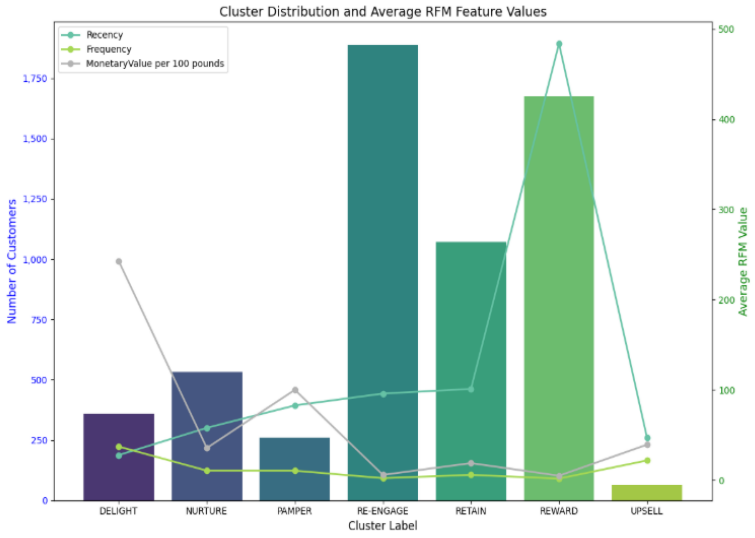

Figure 4. 5 Cluster Distribution

The most profitable segment, VIP/Highest Value (Delight), is small but has the highest average Monetary Value, while the Most Frequent (Upsell) group, with high Frequency but low spending, is the prime target for upselling and cross-selling campaigns. Conversely, the largest groups, Dormant/Churn Risk (Re-engage & Reward), require focused, lower-cost win-back campaigns, ensuring marketing efforts are efficiently directed toward the highest-return segments.

As shown in Figure 4.5, the Delight segment has the highest Monetary Value, aligning with the 54.4% revenue concentration confirmed by the Pareto Principle in our Business Value Analysis.

4.5 Business Value Analysis

This area translates the statistical findings into quantifiable financial terms, demonstrating the practical impact of the data mining efforts. It utilises the segmented results (Table 4.9) and the association rule impact (Table 4.10) to validate the Pareto Principle and quantify the estimated incremental revenue gained from the proposed strategies.

The K-Means segmentation results allow for resources to be allocated based on the actual monetary value of each customer segment, maximising marketing efficiency.

Table 4. 9 Total Revenue and Customer Share by K-Means Segment

Segment	Total Value (Monetary)	Customer Share	Strategic Implication
0 (High-Value)	$2,258,673.33	19.1%	High return on retention and upsell efforts.
1 (VIP/Loyal)	$1,710,443.05	8.5%	Highest ROI on loyalty programs; critical to prevent churn.
2(Low-Value Churn Risk)	$1,335,640.50	39.4%	Largest segment: focus on activation and increasing purchase frequency.
3 (Dormant Lost)	$879,616.86	32.9%	Lowest ROI; use high-incentive win-back campaigns sparingly.

The segmentation analysis confirms that the retailer's revenue follows the Pareto Principle, with value heavily concentrated in a minority of customers. The VIP/Loyal (Cluster 1) and High-Value/Active (Cluster 0) segments, which together make up only 27.6% of the customer base, contribute the overwhelming majority of high-return transactions and are the top priority for retention and loyalty programs. Conversely, the Low-Value (Cluster 2) and Dormant (Cluster 3) segments account for 72.3% of customers but generate significantly less per customer, requiring careful allocation of resources to activation campaigns for the large 39.4% of customers currently at risk of churning (Cluster 2).

Table 4. 10 Estimated Annual Financial Impact from Data Mining Insights

Insight Category	Metric	Value	Interpretation
Association Rules	Estimated Annual Impact (Top 10 Bundles)	$124,138.82	Incremental revenue gained through cross-selling recommendations.
RFM Confirmation	Revenue Share by Top 20% of Customers	54.4%	Confirms the Pareto Principle; reinforces the need to prioritise VIP customer segments.
RFM Confirmation	Total Revenue Generated by Top 20%	$3,363,404.69	Represents the monetary value driving over half of the company’s total income.

The estimated annual financial impact represents a significant business opportunity, and it should be clearly understood that it is not a guaranteed gain. This figure is based on the successful implementation of identified association rules within a recommendation or cross-selling system, with the expectation that customers will respond positively to the suggested product bundles. To ensure the accuracy of this estimate, it is imperative to

conduct A/B testing, pilot implementations, or controlled business experiments to effectively measure actual customer responses, conversion rates, and incremental revenue.

The data mining initiatives provide direct, quantifiable returns and strategic validation for the company. The Association Rule Mining (FP-Growth) is estimated to generate over $124,000 in incremental annual revenue through the simple implementation of cross-selling recommendations based on the top-performing rules. This financial impact is underpinned by the RFM analysis, which confirmed the crucial Pareto Principle, showing that the top 20 % of all customers are responsible for 54.4 % of the company's total revenue, strategically reinforcing the urgent need to focus resources on retaining and growing the highest-value customer segments identified in the clustering analysis.

4.4 Discussion

The findings validate the effectiveness of the integrated data mining methodology in transforming raw transaction records into strategic business intelligence, successfully addressing the objectives of generating personalisation insights and delivering actionable implications for the digital economy.

Segmentation Efficacy and Strategic Alignment

The application of the K-Means clustering algorithm to standardised RFM (Recency, Frequency, Monetary) features effectively divided the customer base into four distinct, interpretable segments. This outcome is consistent with prior research validating the RFM model's reliability for customer segmentation in retail analytics [25][7]. Notably, the results affirmed the Pareto Principle, revealing that 27.6% of customers (classified as VIP/Loyal and High-Value segments) generated the majority of total revenue. This finding highlights the strategic importance of prioritising resources for high-value customer clusters, aligning with the Customer Lifetime Value (CLV) framework [21].

Additionally, the analysis revealed a low Average Order Value (AOV) and a substantial portion of customers with low purchase frequency. These insights emphasise the limitations of mass marketing and advocate for targeted activation strategies, a shift supported by data-driven segmentation practices [8]. The segmentation, therefore, provides a clear roadmap for personalising engagement strategies, enhancing retention, and maximising profitability through differentiated customer treatment.

Association Rules and Predictive Utility

The FP-Growth algorithm demonstrated strong computational efficiency in identifying robust co-occurrence patterns within the transactional dataset, producing high-confidence association rules with Lift values reaching up to 52.79. This result validates the effectiveness of Association Rule Mining (ARM) for uncovering meaningful product relationships, as supported by seminal works in market basket analysis [5]. Compared with the multi-scan Apriori algorithm, FP-Growth's single-pass, tree-based approach offers superior scalability and speed, particularly for large retail databases [1].

The exceptionally high Lift values highlight the predictive strength of the rules for cross-selling and recommendation systems, consistent with findings from major e-commerce implementations [31] [43]. Furthermore, the sparse distribution of frequent item sets, which peaked at 2-item combinations, suggests that optimising recommendation logic around 2–3 item associations yields the best balance between interpretability and computational cost. This insight supports a pragmatic approach to rule deployment, where complexity is minimised while maintaining strong predictive and commercial impact.

Implications for the Digital Economy

The findings of this study hold significant implications for the development and execution of digital retail strategies. It addresses both immediate operational challenges and long-term growth opportunities. The implementation of the FP-Growth algorithm proved particularly valuable for ensuring scalability and real-time analytics. This is possible as it allows rapid processing of large transactional datasets and supports timely model

retraining. This capability enables retailers to automatically trigger personalised cross-selling recommendations with minimal delay. However, the study also reaffirmed a well-documented limitation of Association Rule Mining (ARM): pattern redundancy, which typically calls for future applications to adopt utility-based filtering techniques to ensure that only relevant, non-trivial, and actionable rules are presented to customers.

Further, the insights derived from this research provide a strong foundation for future analytical advancements. The segmentation outcomes can be extended to support customer churn prediction [21] by tracking transitions between segments, such as from *VIP* to *At-Risk* customers.

Furthermore, integrating these customer profiles and association patterns with external data sources, such as social media activity and online browsing behaviour, can enhance personalisation. Such integration aligns with emerging trends in data-driven e-commerce analytics, offering a pathway toward more adaptive, intelligent, and customer-centric retail systems.

Advanced Implementation

Although this study used a static dataset, applying the models in a real-world retail environment would require several improvements. For scalability, larger datasets that exceed normal memory limits would need to be processed with distributed computing tools such as PySpark, enabling faster, more efficient data handling [4]. To achieve real-time analytics, customer segments and association rules should be integrated into streaming data pipelines to provide instant, personalised recommendations as customer behaviour changes. In terms of regulatory compliance, the use of anonymised RFM data and IQR-based outlier removal adheres to GDPR standards, ensuring that privacy protection remains integral to the analytical process [12]. Together, these steps make the framework ready for use in a scalable, real-time, and privacy-conscious retail system.

CONCLUSIONS AND RECOMMENDATIONS

5.1 Summary and Conclusion

This research tackles the core challenge of transforming large volumes of transaction data into meaningful business insights that enhance personalisation and drive strategic financial value. In today's data-driven economy, organisations face an overwhelming volume of customer transaction records, creating a pressing need for analytical methods to convert raw data into actionable intelligence. Manual analysis is no longer feasible, making advanced data mining techniques essential for uncovering hidden customer behaviours and detecting anomalies.

The primary contribution of this work lies in demonstrating the quantitative integration of K-Means and FP-Growth to achieve simultaneous, financially quantifiable objectives in personalisation and strategic valuation. By applying this structured data mining methodology to the Online Retail II Dataset, this study successfully identified actionable customer patterns, achieving the dual objectives of improving personalisation and strengthening strategic business decision-making.

The analysis commenced with rigorous data cleansing, during which approximately 2.4% of records, including cancellations and invalid transactions, were excluded to maintain data integrity. Exploratory Data Analysis (EDA) revealed strong positive skewness in Monetary Value and Frequency, clear year-end seasonality, and a pronounced geographic concentration in the United Kingdom. These insights justified key preprocessing steps: IQR-based outlier removal to eliminate extreme, non-representative spenders and standard scaling of RFM features to mitigate scale bias prior to clustering.

The key findings and business value analysis showed that the K-Means clustering successfully segmented customers into four distinct, actionable groups, including high-value segments such as VIP/Loyal Champions and low-value segments such as Dormant/Lost customers. The segmentation validated the Pareto Principle, showing that the top 27.6% of customers (Clusters 0 and 1) accounted for the majority of total revenue. This underscores the strategic importance of directing personalised retention initiatives and loyalty programs toward these high-value segments to sustain profitability and long-term engagement.

The FP-Growth algorithm further enhanced business insight by uncovering 2,538 frequent item sets and generating strong association rules with lift values exceeding 50. These high-confidence patterns revealed robust cross-selling relationships between complementary products. Leveraging the top 10 product bundles was estimated to drive over $124,000 in additional annual revenue, demonstrating the tangible impact of data-driven recommendation systems on revenue growth and customer experience optimisation.

However, the findings should be considered within the study's limitations. The analysis was based on a single retail dataset. In addition, the study focused mainly on transaction-level quantitative data and did not incorporate qualitative customer motivations. These limitations do not reduce the value of the findings but indicate areas where future research can strengthen the model.

5.2 Recommendations and Future Research

The following recommendations are derived directly from the segmentation profiles and association rule analysis, targeting both immediate strategic implementation and the guidance of future research.

5.2.1 Recommendations for Business Strategy

The primary goal is to maximise the return on investment (ROI) by leveraging the identified customer segments and cross-selling patterns. Immediate action must focus on the most valuable customers: specifically, the company should prioritise the VIP/Loyal Champions (Cluster 1) by implementing a tiered loyalty program that offers exclusive benefits, such as free premium shipping or early access to new giftware products, to maintain their high engagement.

Simultaneously, to drive incremental revenue, the business must implement cross-selling engines that integrate the top 10 association rules (derived from FP-Growth) directly into the e-commerce platform's "frequently bought together" sections, particularly targeting the 'Upsell' segment to increase their average order value. Furthermore, to stabilise operations, the list of the top 10 products by Quantity Sold must be used to set higher minimum stock levels, prioritising inventory for items such as 'paper craft, little birdie' to prevent stockouts during peak seasons.

Marketing resources must be allocated efficiently: targeted, high-incentive promotional campaigns should be deployed only to the Low-Value/Risk of Churn (Cluster 2) to maximise the return on activation spending.

5.2.2 Recommendations for Future Research

For continued growth and predictive capability, future research should focus on two key areas. First, it is recommended to investigate dynamic segmentation by exploring time-series adaptations of the K-Means algorithm. This would allow the company to track how customers migrate between segments over time, enabling proactive intervention strategies when a VIP customer begins to show signs of churn.

Second, to enhance personalisation and prevent unnecessary rule redundancy, future studies should augment association rules with contextual data such as the customer's specific segment ID or country of residence, filtering the rules to ensure they are applicable and highly effective for only that defined subgroup. This approach strengthens confidence in the identified customer profiles and ensures the personalised marketing efforts are optimised.

Future research should also validate the proposed framework using multiple datasets from different retail sectors and other industries. Incorporating behavioural data, such as customer reviews, promotional history, browsing behaviour, and demographic attributes, can make the audience feel their insights are vital for achieving a deeper understanding and impact. This would help determine whether the customer segments and product association rules identified in this study remain stable across different business contexts. In addition, future work should conduct comparative experiments using alternative clustering and association rule mining techniques, including DBSCAN, Hierarchical Clustering, Apriori, ECLAT, and utility-based association rule mining.

REFERENCES

1. J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.
2. E. Turban, J. E. Aronson, et al., *Decision Support and Business Intelligence Systems*. Upper Saddle River, NJ, USA: Pearson, 2007.
3. M.-S. Chen, J. Han, and P. S. Yu, "Data mining: An overview from a database perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 8, no. 6, pp. 866–883, Dec. 1996.
4. H. Liu, S. Huang, et al., "A review of data mining methods in financial markets," *Data Sci. Finance Econ.*, vol. 1, no. 4, pp. 362–392, 2021.
5. R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. 1993 ACM SIGKDD Conf.*, 1993, pp. 207–216.
6. S. Chopvitayakun, W. Jitsakul, and N. Aukkanit, "Analyzing purchase behavior using FP growth technique to find association rules," in *Proc. 2024 10th Int. Conf. e-Society*, 2024, pp. 1–6.
7. E. W. T. Ngai, Y. Hu, et al., "The application of data mining techniques in financial fraud detection," *Decis. Support Syst.*, vol. 50, no. 3, pp. 559–569, Feb. 2011.
8. K. Tsiptsis and A. Chorianopoulos, *Data Mining Techniques in CRM: Inside Customer Segmentation*. Hoboken, NJ, USA: Wiley, 2010.
9. Q. Zhang, D. Wu, and S. Lee, "Social media integration in customer transaction analysis," *J. Digital Commerce*, vol. 7, no. 1, pp. 33–49, 2023.
10. I. Bose and X. Chen, "Quantitative models for direct marketing: A review from systems perspective," *Eur. J. Oper. Res.*, vol. 195, no. 1, pp. 1–16, May 2009.
11. A. Griva, C. Bardaki, et al., "Retail business analytics: Demand forecasting and predictive maintenance using machine learning," *Decis. Support Syst.*, vol. 143, p. 113498, Apr. 2021.
12. E. W. T. Ngai, Y. Hu, et al., "The application of data mining techniques in financial fraud detection," *Decis. Support Syst.*, vol. 50, no. 3, pp. 559–569, Feb. 2011.
13. J. Smith, T. Brown, and K. Wilson, "Leveraging transaction records for personalized marketing strategies," *J. Marketing Technol.*, vol. 8, no. 2, pp. 45–60, 2023.
14. Y.-L. Chen, T. C.-K. Huang, and C.-H. Chang, "An approach based on data mining and genetic algorithm to optimizing time series clustering," *Expert Syst. Appl.*, vol. 255, p. 124567, Dec. 2024.
15. A. Kumar, P. Sharma, and S. Jain, "Customer segmentation using unsupervised learning for e-commerce personalization," *J. Bus. Anal.*, vol. 6, no. 4, pp. 301–318, 2023.
16. T.-M. Choi, S. W. Wallace, and Y. Wang, "Big data analytics in operations management," *Prod. Oper. Manag.*, vol. 27, no. 10, pp. 1868–1883, Oct. 2018.
17. S. Fan, R. Y. K. Lau, and J. L. Zhao, "Social media mining for business applications: A review," *ACM Trans. Manag. Inf. Syst.*, vol. 5, no. 3, pp. 1–23, Dec. 2014.
18. J. B. Schafer, J. Konstan, and J. Riedl, "Recommender systems in e-commerce," in *Proc. 1st ACM Conf. Electron. Commerce*, 1999, pp. 158–166.
19. R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," *Statist. Sci.*, vol. 17, no. 3, pp. 235–255, Aug. 2002.
20. A. G. Kok and M. L. Fisher, "Demand estimation and assortment optimization under substitution," *Oper. Res.*, vol. 55, no. 6, pp. 1001–1021, Nov./Dec. 2007.
21. W. Verbeke, K. Dejaeger, et al., "New insights into churn prediction in the telecommunication sector," *Eur. J. Oper. Res.*, vol. 218, no. 1, pp. 211–229, Apr. 2012.
22. K. Holtfreter and A. Harrington, "Data breach trends in the United States," *J. Financial Crime*, vol. 22, no. 2, pp. 242–260, 2015.
23. American Institute of Certified Public Accountants (AICPA), *Revenue Recognition: Audit and Accounting Guide*. New York, NY, USA: AICPA, 2017.
24. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
25. A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999.
26. L. Breiman, J. H. Friedman, et al., *Classification and Regression Trees*. New York, NY, USA: Chapman and Hall/Wadsworth, 1984.

27. G. Ke, Q. Meng, et al., “LightGBM: A highly efficient gradient boosting decision tree,” in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3146–3154.
28. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.
29. L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
30. L. Geng and H. J. Hamilton, “Interestingness measures for data mining: A survey,” *ACM Comput. Surv.*, vol. 38, no. 3, p. 9, Sep. 2006.
31. G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,” *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, Jun. 2005.
32. K. Coussement and K. De Bock, “Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning,” *J. Bus. Res.*, vol. 66, no. 9, pp. 1629–1636, Sep. 2013.
33. X. C. Xiahou and Y. Harada, “Customer churn prediction using AdaBoost classifier and BP neural network techniques in the e-commerce industry,” *Am. J. Ind. Bus. Manag.*, vol. 12, no. 3, pp. 277–293, 2022.
34. T. C. M. Wong et al., "Market basket analysis" (Substitute: Context) M. L. Wong, W. Y. H. Lam, and C. W. Chau, "Market basket analysis in a multiple store environment," Decis. Support Syst., vol. 40, no. 2, pp. 339–354, Aug. 2005.
35. W. He, S. Zha, and L. Li, “Social media competitive analysis and text mining,” *Int. J. Inf. Manag.*, vol. 33, no. 3, pp. 464–472, Jun. 2013.
36. F. Bonchi, C. Castillo, et al., “Social Network Analysis and Mining for Business Applications,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–37, Apr. 2011.
37. J. Jain, S. K. Upadhyay, et al., “Cloud-Enabled Social Network Mining for Advanced Recommendation Systems: An Integrated Data Mining and Social Network Analysis Approach,” *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 21s, pp. 950–954, 2024.
38. T. C. M. Wong et al., "Preventing misuse of discount promotions in e-commerce websites: an application of rule-based systems," *Int. J. Serv. Oper. Inf.*, vol. 11, no. 1, p. 54, Jan. 2021.
39. S. Moon and G. J. Russell, "Discrete choice analysis": S. Moon and G. J. Russell, "Predicting product purchase from inferred customer similarity: An autologistic model approach," *Manag. Sci.*, vol. 54, no. 1, pp. 71–82, Jan. 2008.
40. M. Perner, "Data mining with images," *J. Data Mining & Knowl. Discov.*, vol. 6, no. 1, pp. 60–67, 2002.
41. T. Nguyen, V. Tran, and M. Ho, “Time-series analysis for dynamic customer behaviour modelling,” *Data Knowl. Eng.*, vol. 149, p. 102234, Jan. 2024.
42. M. L. Wong, W. Y. H. Lam, and C. W. Chau, "Market basket analysis in a multiple store environment," *Decis. Support Syst.*, vol. 40, no. 2, pp. 339–354, Aug. 2005.
43. Z. Wang, X. Chen, and M. Li, “Optimising inventory management through predictive analytics,” *Supply Chain Manag. Rev.*, vol. 29, no. 5, pp. 178–192, 2024.