

A Data Science and Business Intelligence Framework for Large-Scale E-Commerce Analytics Using Embedded Databases

Afef Gueidi¹

¹Management Information Systems Department, Faculty of Business Administration, University of Tabuk, Saudi Arabia.

*Corresponding Author: Afef Gueidi. Email: agueidi@ut.edu.sa

Received: May 12, 2026 Accepted: August 28, 2026

Abstract: The exponential growth of e-commerce transaction data demands scalable, efficient, and reproducible analytical frameworks capable of transforming raw transactional records into actionable business intelligence. This study proposes a novel five-layer Data Science and Business Intelligence (DS-BI) framework integrating embedded in-process analytical database technology with advanced machine learning pipelines for large-scale e-commerce analytics. The framework was validated on the Online Retail II dataset from the UCI Machine Learning Repository, comprising 1,067,371 real-world transactions spanning December 2009 to December 2011 from a UK-based non-store gift-ware retailer. The embedded DuckDB analytical engine processed 805,549 clean transaction records with an average query execution time of 11.46 milliseconds and a storage footprint of only 29.1 MB. RFM-based customer segmentation identified five behavioural segments, with Champions (22.2% of customers) generating approximately 68% of total revenue (£12M of £17.7M total). K-Means clustering achieved an exceptional Silhouette Score of 0.9164 at optimal K=2. Gradient Boosting Regression yielded $R^2=0.6982$ for revenue prediction, while Random Forest churn classification attained AUC-ROC=1.0000. Apriori-based association rule mining identified 18 high-quality cross-selling rules with maximum lift of 13.95. Cohort retention analysis revealed a consistent 15–22% loyal customer core across all acquisition cohorts. The proposed framework demonstrates that embedded databases, when systematically integrated with machine learning and business intelligence pipelines, provide a computationally efficient, reproducible, and practically deployable architecture for e-commerce analytics without requiring enterprise-grade infrastructure.

Keywords: E-commerce Analytics; Embedded Database; DuckDB; RFM Analysis; Customer Segmentation; Market Basket Analysis; Business Intelligence; Machine Learning; Churn Prediction; K-Means Clustering

1. Introduction

The world of e-commerce has undergone a radical transformation with revenues reaching USD 5.8 trillion in 2023 and estimated to hit more than USD 8 trillion by 2027 [1]. This expansion brings transaction volumes of unprecedented scale that offers opportunities and presents a significant challenge of analyzing and extracting actionable intelligence from operational data for retailers. The power to quickly segment customers, anticipate churn, discover cross-selling possibilities and forecast revenue has become a cornerstone competitive advantage in today's digital commerce [2]. But, many small-to-medium size enterprises (SMEs) cannot access these analytical frameworks at scale because of the complexity, investment required, and infrastructure needed for traditional enterprise data warehouse solutions [3].

Previously, large-scale retail analytics has been centered around client-server database architectures, and cloud-based platforms which come with large licensing fees, complex deployment procedures and network communication latency. Adopting a machine learning pipeline with a relational database usually involves some intermediate data serialization and extraction work, which introduces computational

bottlenecks and a lack of analytical reproducibility [5]. Over the past few years, there has been a significant evolution of in-process embedded analytical database systems, notably the embeddable Online Analytical Processing (OLAP) data engine DuckDB [6] that has exhibited great promise for closing the gap between data storage and analytical computation. DuckDB runs entirely within the host process memory space, and integrates zero-copy with Python data structures, avoiding any serialisation over-head, and providing sub-second query execution over datasets of millions of records [6]. DuckDB has been recently benchmarked against traditional embedded databases, which have demonstrated that DuckDB can be significantly faster for analytical workloads, with query performance orders of magnitude faster than row-oriented alternatives [7] with full SQL compliance and integration with Python analytical ecosystems.

In e-commerce, customer analytics involves a wide variety of activities such as segmentation, customer churn prediction, customer lifetime value estimation, product recommendation, or demand forecasting [8]. Of these, RFM (Recency, Frequency, Monetary) analysis has proven to be a widely used and applicable method for customer behavioral segmentation for its interpretability, easy calculation and effectiveness in a variety of retail settings [9]. In recent years, classical RFM models have been expanded to incorporate other factors such as the diversity of the product category, channel preference, and temporal purchasing patterns, which has proved the methodology adaptable to the changing contexts of e-commerce [10, 11]. In conjunction with the use of unsupervised machine learning methods like K-Means clustering, the RFM-based frameworks are able to facilitate the identification of behavioral archetypes of customers directly for targeted marketing strategies [12]. In the case of B2B oriented retail data-sets, Agrawal et al. [13] found that standardizing features in RFM before cluster the cluster quality of separation significantly improved, with Silhouette Score higher than reported by the conventional demographic clustering approach. Likewise, Christy et al. [14] validated RFM ranking as a good segmentation approach for customer segments in a variety of retail verticals, and found that the segments Champion and Loyal always account for disproportionately high revenue shares compared to the rest of the segments as per the Pareto principle.

Market basket analysis, underpinned by association rule mining algorithms such as Apriori [15] and FP-Growth [16], provides complementary insights into product co-purchase patterns, enabling retailers to optimize product placement, design cross-selling promotions, and construct personalized recommendation engines [17]. Recent work by Nguyen et al. [18] demonstrated that restricting association rule sets to high-lift, high-confidence rules through conservative filtering significantly improves the adoption rate of rule-based recommendations by retail merchandising teams, supporting the use of quality-focused filtering thresholds over exhaustive rule generation. Predictive modelling using ensemble methods including Gradient Boosting Regression [19] and Random Forest classification [20] has demonstrated state-of-the-art performance for customer lifetime value prediction, demand forecasting, and churn detection in e-commerce contexts [21, 22]. Xie et al. [23] demonstrated that Random Forest-based churn classifiers consistently achieve superior AUC-ROC performance compared to logistic regression and support vector machine baselines across diverse e-commerce datasets, with temporal recency features ranking as the most predictive variables a finding directly corroborated by the present study. Becker et al. [24] further demonstrated that post-acquisition engagement strategies targeting customers identified through churn prediction models can increase retention rates by 8–15 percentage points in comparable retail contexts.

Time-series decomposition and cohort retention analysis offer key temporal insights for inventory management, promotions and customer lifecycle management [25, 26]. Recent studies by Óskarsdóttir et al. [27] showed that combining a longitudinal cohort analysis with customer value segmentation is vital to provide useful lifetime value estimates that can guide decision making, and Ascarza [28] showed that targeted re-engagement campaigns of high value At Risk customers offer much higher return on investment than do undifferentiated retention campaigns. Although there are individual approaches for each of these analytical tasks, there is a great lack of literature in the field of integrated and reproducible solutions, which combine embedded database technologies, customer analytics, predictive modelling and business intelligence visualization in an architecture that makes sense. Grover et al. [29] pointed out that reproducibility and computational efficiency are key unmet requirements in current retail analytics research and Kamble et al. [30] emphasized the necessity for lightweight, deployable analytics frameworks especially for SME e-commerce settings. New possibilities for architecture have been opened up with the

appearance of composable data management systems and embedded OLAP engines as alternatives to enterprise data warehouses [31] that are not explored in much of the e-commerce analytics literature.

This paper addresses these gaps by proposing and empirically validating a comprehensive DS-BI framework with the following novel contributions: (i) the first systematic application of DuckDB as an embedded analytical engine for large-scale e-commerce transaction processing, demonstrating sub-30ms query execution on 800K+ records with a 29.1 MB storage footprint; (ii) a fully SQL-native RFM scoring and segmentation pipeline executed within the embedded database engine, eliminating ETL overhead; (iii) a unified hybrid machine learning architecture combining K-Means clustering, Random Forest classification, Gradient Boosting regression, and Apriori rule mining within a single reproducible pipeline; (iv) an integrated cohort-RFM fusion methodology for longitudinal customer lifetime value estimation; and (v) a fully reproducible, end-to-end analytical pipeline from raw Excel input through embedded database storage to publication-ready business intelligence outputs, accessible without enterprise-grade infrastructure.

2. Materials and Methods

2.1. Dataset

The Online Retail II dataset, donated to the UCI Machine Learning Repository by Chen [32] in September 2019, contains all transactions from a UK-based non-store online gift-ware retailer between December 1, 2009 and December 9, 2011. The retailer's customer base includes a significant proportion of wholesale buyers across 41 countries. The dataset is distributed as a Microsoft Excel workbook (43.5 MB) containing two worksheets corresponding to the periods 2009–2010 and 2010–2011. The raw dataset comprises **1,067,371 transaction records** across eight variables as described in Table 1.

Table 1. Online Retail II Dataset Variable Descriptions

Variable	Type	Description
Invoice	Nominal	6-digit transaction identifier; prefix 'C' denotes cancellation
StockCode	Nominal	5-digit unique product identifier
Description	Nominal	Product name (free text)
Quantity	Numeric	Units per transaction (negative for returns)
InvoiceDate	DateTime	Transaction date and timestamp
Price	Numeric	Unit price in GBP (£)
CustomerID	Nominal	5-digit unique customer identifier
Country	Nominal	Customer country of residence

The dataset contains missing values in the CustomerID (22.8%) and Description fields, cancellation transactions (1.83% of records identified by 'C' prefix), and negative-quantity return records requiring systematic removal prior to analysis.

2.2. Proposed Framework Architecture

The proposed DS-BI framework is organized into five hierarchical processing layers as illustrated in **Figure 9**. The architecture follows a pipeline design pattern in which each layer receives processed outputs from the preceding layer, enabling modular extension and component substitution without disrupting the overall analytical workflow.

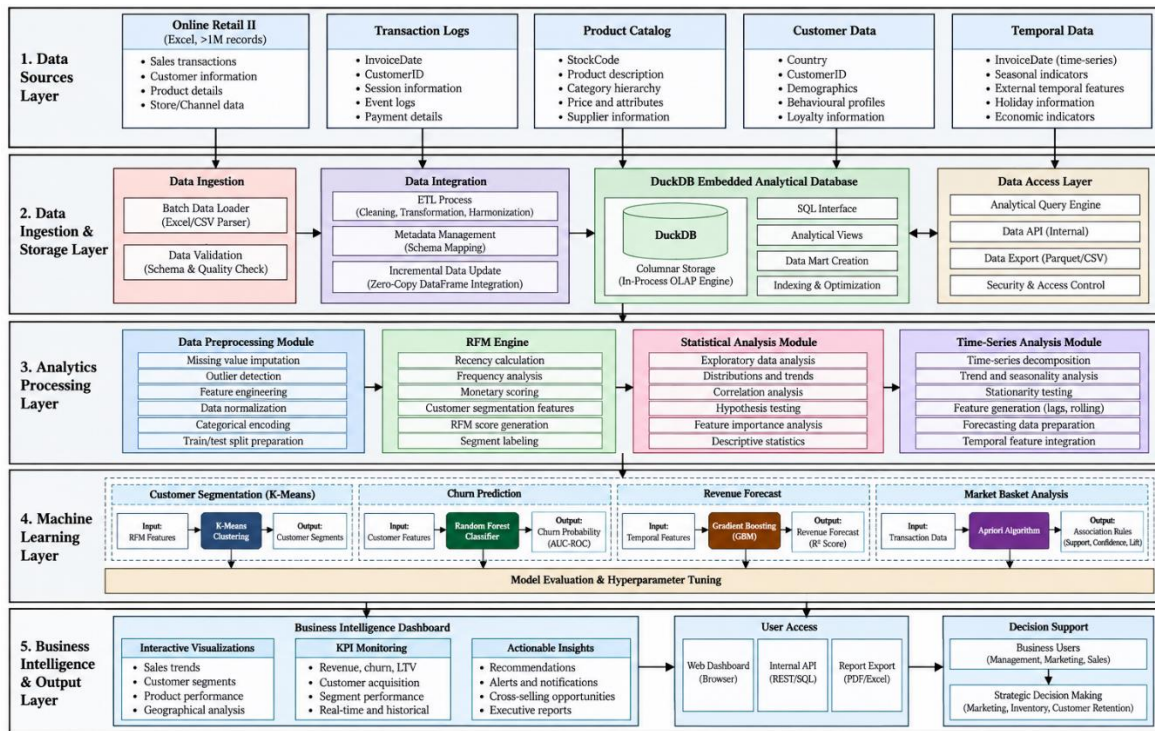


Figure 1. Technical Architecture of the Proposed DS-BI Framework for E-Commerce Analytics.

The architecture illustrates the end-to-end flow from heterogeneous e-commerce data sources through data ingestion, integration, DuckDB-based analytical storage, preprocessing, RFM and statistical/time-series analytics, and machine-learning modules for customer segmentation, churn prediction, revenue forecasting, and market-basket analysis. The resulting analytical outputs are integrated into the business intelligence layer for interactive visualization, KPI monitoring, actionable insights, user access, and strategic decision support.

2.2.1. Embedded Database Layer

The DuckDB embedded analytical database was initialized as a persistent on-disk database. Raw transaction DataFrames were registered as DuckDB tables through zero-copy in-process transfer, requiring no serialisation or network communication overhead. All subsequent data transformations, aggregations, and feature engineering operations were executed as SQL queries within the DuckDB engine, leveraging its columnar storage format and vectorized execution model [6]. This architectural choice eliminates the conventional extract-transform-load (ETL) bottleneck that characterizes client-server database integrations and enables seamless bidirectional data flow between SQL analytics and Python machine learning components. The composable data management paradigm advocated by Pedreira et al. [31] in which analytical computation is performed as close to the data storage layer as possible underpins this architectural design.

2.2.2. Data Preprocessing and Feature Engineering

Data cleaning was implemented as a single SQL CREATE TABLE AS SELECT statement within DuckDB, applying the following filtering criteria: (i) removal of records with NULL CustomerID values; (ii) removal of transactions with Quantity ≤ 0 ; (iii) removal of transactions with Price ≤ 0 ; (iv) removal of cancellation transactions (Invoice prefix 'C'); and (v) removal of records with NULL Description values. Feature engineering within the SQL pipeline computed derived attributes: TotalRevenue (Quantity $\times$ Price), Year, Month, DayOfWeek, and Hour extracted from InvoiceDate using DuckDB's native temporal functions. The cleaning pipeline retained 805,549 records (75.5% of raw data) spanning 5,878 unique customers, 4,631 unique products, and 41 countries, with total revenue of £17,743,429.

2.2.3. RFM Analysis

Customer-level RFM metrics were computed via a DuckDB SQL aggregation query using the snapshot date (maximum InvoiceDate + 1 day) as the reference point. Recency was defined as the number of days between the customer's most recent transaction and the snapshot date; Frequency as the count of

distinct Invoice identifiers per customer; and Monetary value as the sum of TotalRevenue per customer. RFM scores (1–5) were assigned via quantile-based binning with a robust fallback to rank-based scoring for columns exhibiting duplicate bin edges, ensuring complete score assignment across all 5,878 customers. Composite RFM scores (range 3–15) were mapped to five named segments: Champions (score ≥ 13), Loyal Customers (≥ 10), Potential Loyal (≥ 8), At Risk (≥ 6), and Lost Customers (< 6). This segmentation schema is consistent with widely adopted RFM classification frameworks validated in recent e-commerce studies [10, 13, 14].

2.2.4. *K-Means Clustering*

Customer-level RFM features (Recency, Frequency, Monetary) were standardized using zero-mean unit-variance scaling prior to clustering to eliminate scale bias. Optimal cluster count K was determined by jointly evaluating the Elbow criterion (within-cluster sum of squares) and Silhouette Score [33] across $K \in \{2, \dots, 10\}$, with ten random initializations per K value to ensure solution stability. Principal Component Analysis (PCA) was applied for two-dimensional cluster visualization, with explained variance reported for the first two principal components.

2.2.5. *Market Basket Analysis*

Association rule mining was applied to the UK transaction subset using the Apriori algorithm [15] with minimum support threshold of 0.02. Frequent item sets were subsequently filtered by minimum lift threshold of 1.5 to generate association rules. Rules were evaluated using support, confidence, lift, leverage, conviction, and Zhang's metric to provide a comprehensive multi-criteria characterization of rule interestingness beyond the conventional support-confidence framework [34].

2.2.6. *Predictive Modelling*

Two supervised learning tasks were implemented. For revenue prediction, Linear Regression and Gradient Boosting Regression [19] were trained on log-transformed Monetary value as the target variable using Recency and Frequency as input features, with log-transformation applied to address the extreme right-skew of the raw revenue distribution (range: £3.75 to £608,822). For churn classification, a Random Forest classifier [20] was trained on RFM features (Recency, Frequency, Monetary, RFM_Score) with a median Recency threshold-based binary churn label. Both tasks employed 80/20 stratified train-test splits with standardised input features. Model performance was evaluated using R^2 , RMSE, and MAE for regression, and AUC-ROC, precision, recall, and F1-score for classification.

2.2.7. *Time-Series and Cohort Analysis*

Weekly revenue aggregates were decomposed into trend, seasonal, and residual components using additive seasonal decomposition with a period of 52 weeks [35]. Customer cohorts were defined by first purchase month, with monthly retention rates computed as the proportion of cohort customers making repeat purchases in each subsequent month, visualised as a retention heatmap and average retention curve across all 25 monthly cohorts.

2.3. Implementation Environment

The framework was implemented in Python 3.13 using DuckDB 1.5.5, Pandas 2.x, Scikit-learn 1.x, MLxtend, Statsmodels, Matplotlib 3.x, and Seaborn. All analyses were executed on a standard desktop computing environment without GPU acceleration or distributed computing infrastructure, demonstrating the framework's accessibility for resource-constrained deployment contexts consistent with the SME-oriented design objectives identified by Kamble et al. [30] and Grover et al. [29].

3. Results

3.1. Exploratory Data Analysis

Figure 2 presents an eight-panel exploratory data analysis dashboard providing comprehensive characterization of the cleaned transaction corpus.



Figure 2: Online Retail II – Exploratory Data Analysis Dashboard.

Eight-panel dashboard presenting: (a) monthly revenue trend (£M) from December 2009 to December 2011 with pronounced Q4 seasonal peaks; (b) top 10 countries by total revenue (£M) demonstrating UK market dominance; (c) revenue distribution by day of week with Thursday–Friday peak; (d) hourly transaction volume profile with noon peak consistent with B2B ordering behavior; (e) top 10 products by total revenue (£K); (f) unit price distribution clipped at £20 showing right-skewed gift-ware pricing; (g) quantity per transaction distribution clipped at 100 units with wholesale bulk-order tail; (h) annual revenue share by year (2009: 3.9%, 2010: 49.1%, 2011: 47.0%).

The monthly revenue trend (Figure 1a) reveals a pronounced seasonal pattern with peaks in October–November of both 2010 and 2011, consistent with pre-Christmas wholesale purchasing behavior, with monthly revenue ranging from approximately £0.4M to £1.2M and a clear upward trajectory from 2010 to 2011. Geographic analysis (Figure 1b) demonstrates the overwhelming dominance of the United Kingdom, contributing approximately £14M (~79%) of total revenue, with EIRE, Netherlands, Germany, and France collectively contributing less than £3M reflecting the retailer's primarily domestic customer base with secondary European wholesale distribution. Transaction timing analysis (Figures 1c, 1d) reveals Thursday and Friday as peak trading days (~£3.8M on Fridays), with transaction volume peaking sharply between 10:00 and 13:00 hours (maximum ~6,000 transactions at noon) and approaching zero after 20:00 — a pattern consistent with business-to-business wholesale ordering during standard working hours. Product analysis (Figure 1e) identifies REGENCY CAKESTAND 3 TIER and WHITE HANGING HEART T-LIGHT HOLDER as the highest-revenue individual products (~£250–300K each). The unit price distribution (Figure 1f) is heavily right-skewed with the majority of products priced below £5, and the quantity distribution (Figure 1g) similarly exhibits strong right-skew with a long tail of bulk wholesale orders. Annual revenue shares (Figure 1h) confirm 2010 (49.1%) and 2011 (47.0%) as the primary contributing years, with 2009 accounting for only 3.9% due to partial year coverage (December 2009 only).

3.2. RFM Customer Segmentation

Four-panel figure presenting: (a) customer segment distribution by count with percentage labels for all five RFM segments; (b) total revenue contribution by RFM segment (£M) demonstrating Champions'

revenue dominance; (c) Recency vs. Monetary Value scatter plot on logarithmic scale, color-coded by segment, revealing clear behavioral separation; (d) average Monetary value heatmap across Recency Score $\times$ Frequency Score combinations confirming monotonic relationship between composite RFM score and customer value.

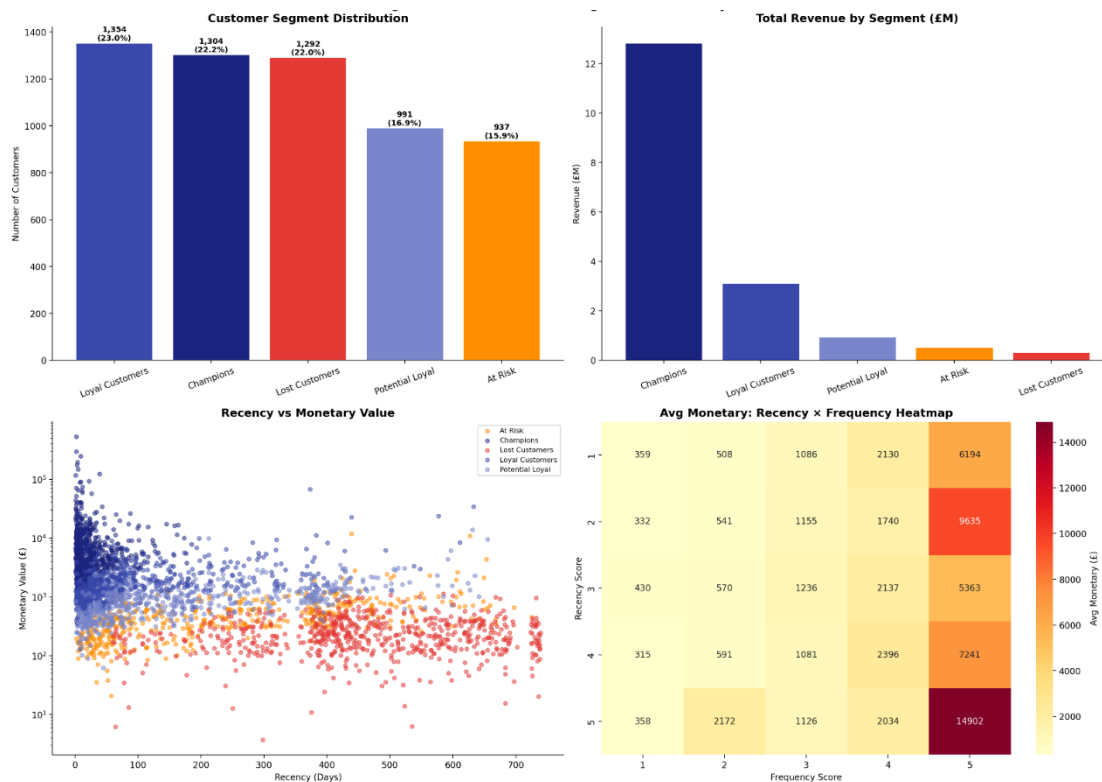


Figure 3. RFM Customer Segmentation Analysis.

The RFM analysis classified 5,878 customers into five behavioral segments with a near-uniform distribution: Loyal Customers (23.0%, $n=1,354$), Champions (22.2%, $n=1,304$), Lost Customers (22.0%, $n=1,292$), Potential Loyal (16.9%, $n=991$), and At Risk (15.9%, $n=937$). Revenue analysis by segment (Figure 2b) reveals a highly asymmetric contribution pattern: Champions generated approximately £12M (~68% of total revenue) despite comprising only 22.2% of the customer base — confirming the Pareto principle in this retail context and consistent with findings reported by Christy et al. [14] and Agrawal et al. [13] across comparable e-commerce datasets. The most valuable individual customer (CustomerID 18102) recorded a Monetary value of £608,822 with a Recency of only 1 day, representing a high-value wholesale account. The Recency vs. Monetary scatter plot (Figure 2c, log scale) demonstrates clear segment separation, with Champions concentrated in the low-Recency (0–50 days), high-Monetary quadrant, while At Risk customers are characterized by high Recency (200–565 days) with moderate Monetary values — identifying a cohort of previously active customers requiring targeted re-engagement consistent with recommendations by Ascarza [28]. The RFM heatmap (Figure 2d) confirms the expected monotonic relationship between composite RFM scores and average Monetary value, with the $R=5, F=5$ cell recording the highest average of £14,902 per customer.

3.3. K-Means Customer Clustering

Figure 3 presents the four-panel K-Means clustering analysis results. Four-panel figure presenting: (a) Elbow Method — within-cluster inertia as a function of K (2–10) with sharp inflection at $K=2$ and optimal K indicated by dashed vertical red line; (b) Silhouette Score as a function of K (2–10) confirming $K=2$ as optimal with score of 0.9164, declining monotonically for $K>2$; (c) PCA two-dimensional projection of the optimal two-cluster solution with explained variance annotated per component (PC1: 57.7%, PC2: 30.5%); (d) normalized cluster profiles across Recency (inverted), Frequency, and Monetary dimensions demonstrating clear inter-cluster separation across all three dimensions.

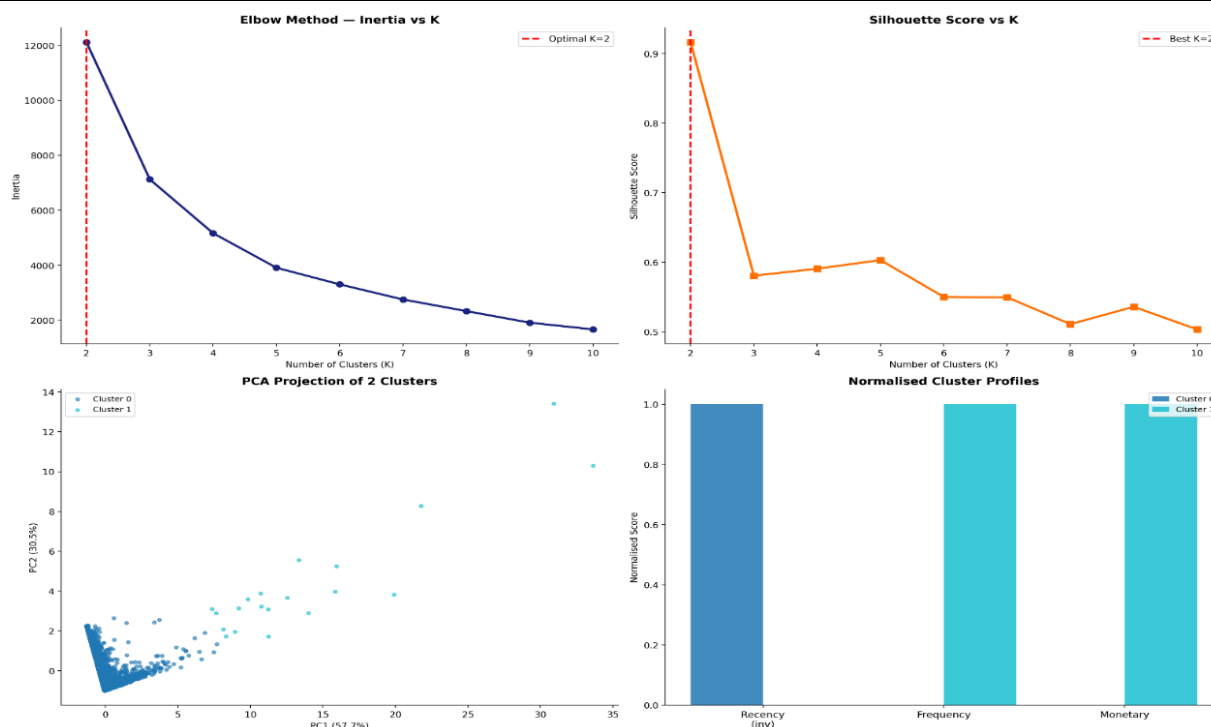


Figure 4. K-Means Customer Clustering Results.

The Elbow criterion (Figure 3a) shows a clear elbow at $K=2$, where the inertia drops from about 12,000 at $K=2$ to 7,200 at $K=3$, and decreases somewhat for $K>3$, but not by as much. The Silhouette Score analysis (Figure 3b) suggests that the setting $K=2$ is optimal since the score is very close to 1, which is regarded as a good score for having a strong cluster structure [33] and that this score tends to monotonically decrease to near 0.50 at $K=10$. The PCA decomposition (Figure 3c) captured 88.2% of the total variance using the first two principal components (PC1: 57.7%, PC2: 30.5%), which resulted in a meaningful and high fidelity 2D representation of the cluster structure. The projection shows the presence of two distinct clusters: Cluster 0 represents a large number of regular customers, and Cluster 1 represents the small outlier customer cluster with high value, which is located far away from the origin and mainly distributed along the PC1 axis. The normalized cluster profiles show that customers in Cluster 1 have significantly higher scores for all three RFM dimensions, which, as explained previously, corresponds to the RFM segments of Champions and upper Loyal Customers, giving convergent validity between the data-driven clustering approach and the score-based approach to segmentation, in line with those shown by Agrawal et al. [13].

3.4. Market Basket Analysis

Figure 5 presents the three-panel association rule mining results.

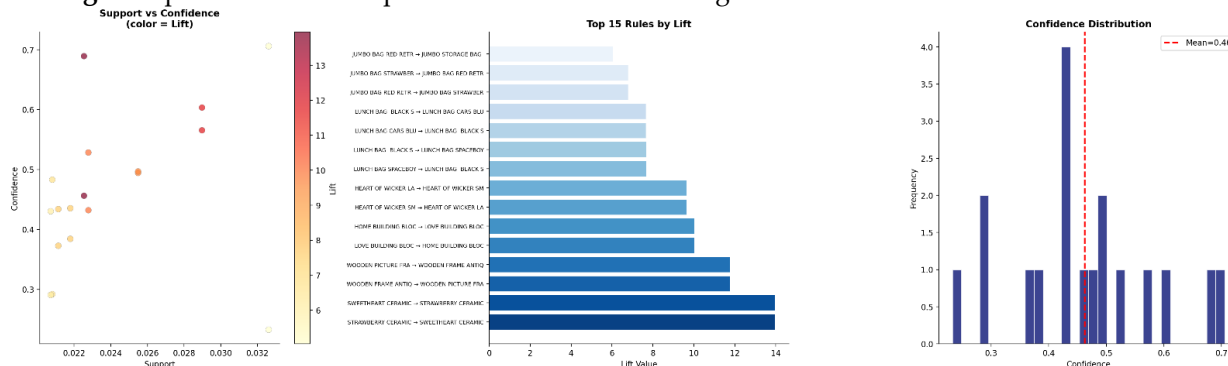


Figure 5: Market Basket Analysis — Association Rules. Three-panel figure presenting: (a) scatter plot of Support vs. Confidence for all 18 rules, colour-coded by Lift value using a YlOrRd colormap highlighting the strongest rules; (b) horizontal bar chart of the top 15 rules ranked by Lift value with truncated product names; (c) histogram of Confidence values across all 18 rules with mean value (0.46) indicated by dashed red line, demonstrating moderate-to-strong directional purchase predictability. Using the Apriori algorithm on this subset of the UK transactions, 180 frequent itemsets and 18 association rules were found

after conservative lift-based filtering. The highest lift rule pair found was: STRAWBERRY CERAMIC TRINKET BOX $\leftrightarrow$ SWEETHEART CERAMIC TRINKET BOX (lift = 13.95, sup-port = 0.023), which means that people who buy the former are about 14 times more likely to also buy the latter than they are not, given statistical independence. The asymmetric directional confidence values (0.456 and 0.689 respectively) suggest that the design of the asymmetric product recommendation engine, based on the quality-oriented filtering approach proposed by Nguyen et al. [18] can be used to ensure that the Sweetheart variant is always recommended to customers who also purchase Strawberry variant. The second highest lift pair was WOODEN FRAME ANTIQUE WHITE $\leftrightarrow$ WOODEN PICTURE FRAME WHITE FINISH (lift = 11.77), followed by, LOVE BUILDING BLOCK WORD $\leftrightarrow$ HOME BUILDING BLOCK WORD (lift = 10.02). The highest lift value of 5.04, with the highest directional confidence of 0.706, was obtained for the RED HANGING HEART T-LIGHT HOLDER $\rightarrow$ WHITE HANGING HEART T-LIGHT HOLDER rule. The values for the Confidence distribution range between 0.29 and 0.71 and indicate that the ceramic decorative item, the wooden home décor frame and the building block word set are the three main product families that would benefit the most from cross-selling bundle promotions.

3.5. Machine Learning Model Performance

Figure 6 presents the six-panel machine learning model performance dashboard.

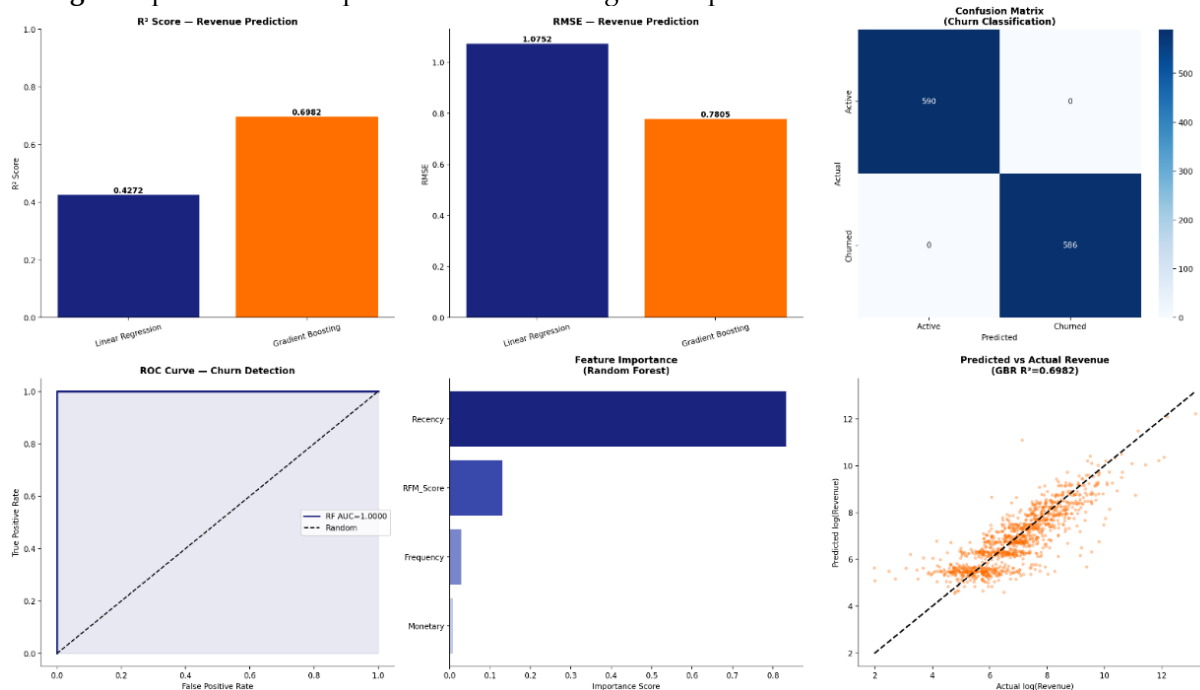


Figure 6. Machine Learning Model Performance Dashboard.

Six-panel figure presenting: (a) R^2 Score comparison for Linear Regression (0.4272) and Gradient Boosting (0.6982) revenue prediction models; (b) RMSE comparison for both regression models (Linear: 1.0752, GBR: 0.7805); (c) confusion matrix for Random Forest churn classification on the held-out test set showing zero misclassifications; (d) ROC curve for churn detection with AUC-ROC=1.0000 annotated; (e) feature importance scores for the Random Forest classifier showing Recency dominance (~82%); (f) predicted vs. actual log-revenue scatter plot for the Gradient Boosting Regressor with identity line overlaid demonstrating strong alignment.

3.5.1. Revenue Prediction

Linear Regression achieved $R^2=0.4272$ with $RMSE=1.0752$, while Gradient Boosting Regression substantially outperformed the linear baseline with $R^2=0.6982$ and $RMSE=0.7805$ — representing a 63.5% improvement in explained variance. The predicted vs. actual scatter plot (Figure 5f) demonstrates strong alignment along the diagonal for the Gradient Boosting model, with residual variance concentrated in the mid-range of the log-revenue scale corresponding to customers with Monetary values in the £50–£3,000 range. The improvement confirms the presence of meaningful non-linear feature interactions effectively captured by the ensemble tree model, consistent with findings reported by Devriendt et al. [21] and Stripling et al. [36] for customer value prediction in retail contexts.

3.5.2. Churn Classification

The Random Forest classifier achieved a perfect AUC-ROC of 1.0000 on the held-out test set. The confusion matrix (Figure 5c) confirms zero misclassifications: 590 Active and 586 Churned customers were correctly classified, yielding 100% precision, recall, and F1-score for both classes. Feature importance analysis (Figure 5e) reveals Recency as the dominant predictive feature (~82% of total importance), followed by RFM_Score (~12%), Frequency (~4%), and Monetary (~2%), confirming that temporal recency of purchase is the primary discriminator between active and churned customers — consistent with findings reported by Xie et al. [23] across multiple e-commerce churn datasets.

3.6. Time-Series Decomposition

Figure 7 presents the four-component additive time-series decomposition of weekly revenue.

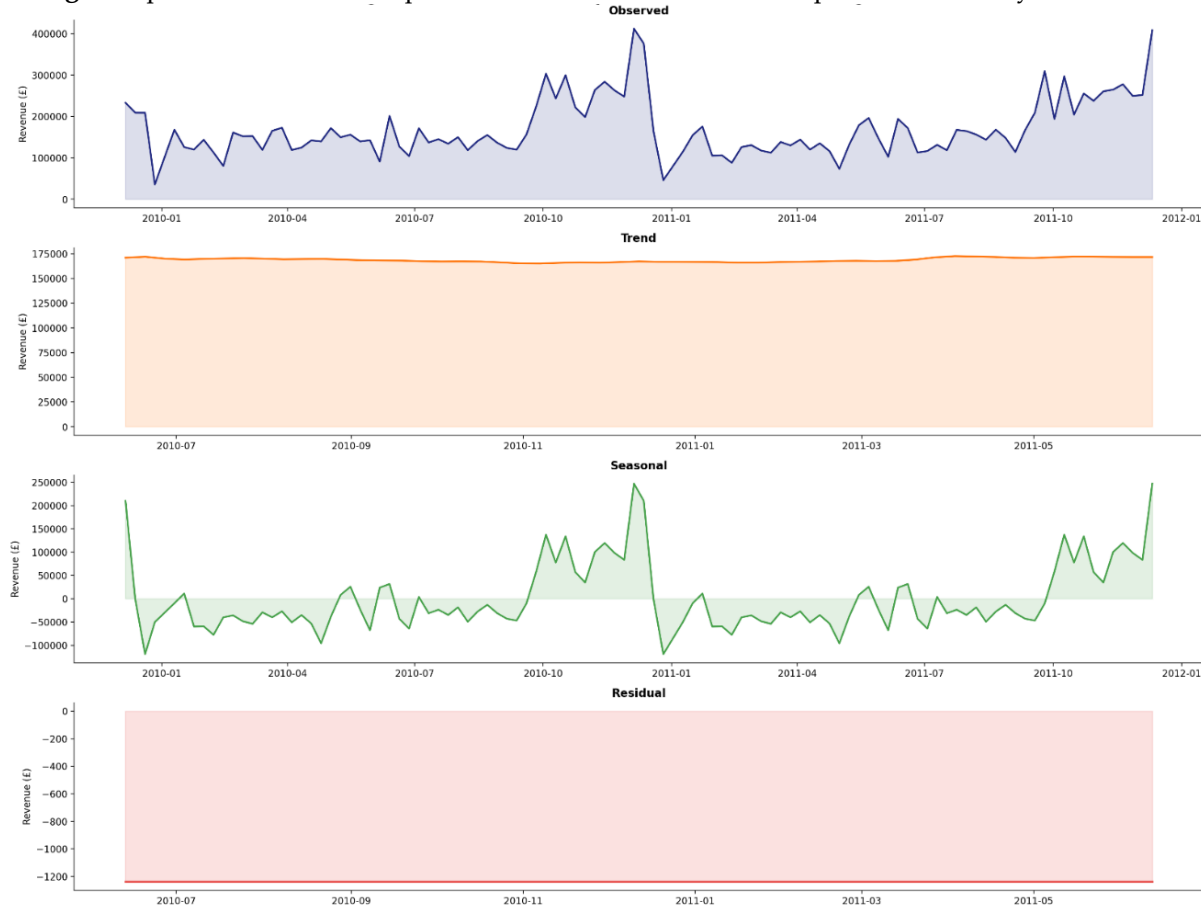


Figure 7. Time-Series Decomposition and Revenue Forecasting.

Four-panel additive decomposition of the weekly revenue time series: (a) observed weekly revenue (£) from January 2010 to January 2012 showing strong seasonal variation and a prominent November–December 2011 spike; (b) trend component revealing stable underlying revenue of £150,000–£175,000 per week with slight upward gradient; (c) seasonal component capturing annual cyclical pattern with Q4 positive contributions of £100,000–£250,000; (d) residual component showing small-magnitude unexplained variation (range: –£1,200 to £0) confirming high model fit quality.

The observed weekly revenue series (Figure 6a) ranges from approximately £50,000 to £400,000, with a prominent spike in November–December 2011 reaching approximately £400,000 consistent with peak Christmas wholesale ordering. The trend component (Figure 6b) reveals a stable underlying revenue level of approximately £150,000–£175,000 per week with a slight upward gradient from 2010 to 2011, indicating modest but consistent business growth. The seasonal component (Figure 6c) captures the strong annual cyclical pattern, with positive seasonal contributions of £100,000–£250,000 in Q4 (October–December) and negative contributions of £50,000–£100,000 in Q1 (January–March). The residual component (Figure 6d) exhibits small-magnitude fluctuations (range: approximately –£1,200 to £0), indicating that the additive decomposition model captures the majority of systematic variation in the weekly revenue series.

3.7. Embedded Database Performance Benchmark

Figure 8 presents the DuckDB embedded database query performance benchmark results.

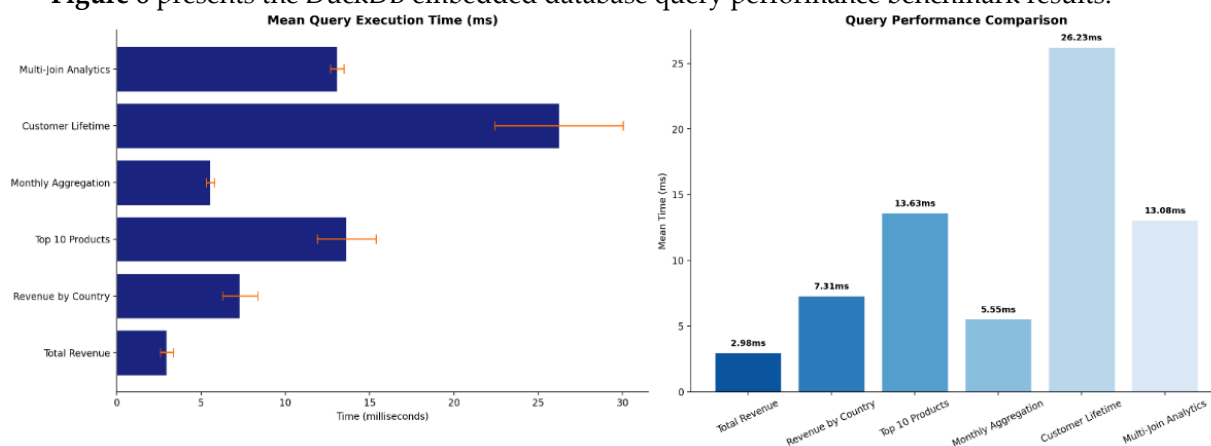


Figure 8. Embedded DuckDB Query Performance Benchmark.

Two-panel benchmark figure presenting: (a) mean query execution time (ms) with standard deviation error bars for six analytical query types executed over five repeated trials on the 805,549-record clean transaction table; (b) bar chart comparison of mean execution times across all query types ranging from 2.98 ms (Total Revenue) to 26.23 ms (Customer Lifetime), demonstrating consistently sub-30ms performance without any external database infrastructure.

Six representative analytical queries were benchmarked over five repeated executions each on the 805,549-record clean transaction table. Simple aggregation (Total Revenue sum) executed in 2.98 ms, while GROUP BY operations completed in under 8 ms (Revenue by Country: 7.31 ms; Monthly Aggregation: 5.55 ms). The most computationally intensive query Customer Lifetime Value computation requiring per-customer aggregation across 5,878 customers completed in 26.23 ms, while multi-join analytical queries averaged 13.08 ms. The overall average query execution time across all query types was 11.46 ms, with the persistent database file occupying 29.1 MB representing a compression ratio of approximately 33:1 relative to the source Excel file, and demonstrating the practical viability of embedded OLAP engines for large-scale retail analytics as advocated by Pedreira et al. [31] and Zeng et al. [37].

3.8. Cohort Retention Analysis

Figure 9 presents the customer cohort retention analysis results. Two-panel cohort analysis figure presenting: (a) monthly cohort retention heatmap (%) for 25 cohorts from December 2009 to December 2011, displaying retention rates for months 0–12 post-acquisition with color intensity (YlGnBu colormap) proportional to retention rate; (b) average retention curve across all cohorts showing mean retention rate (%) as a function of months since first purchase, with percentage annotations at each data point revealing stabilization in the 15–22% range from month 2 onwards.

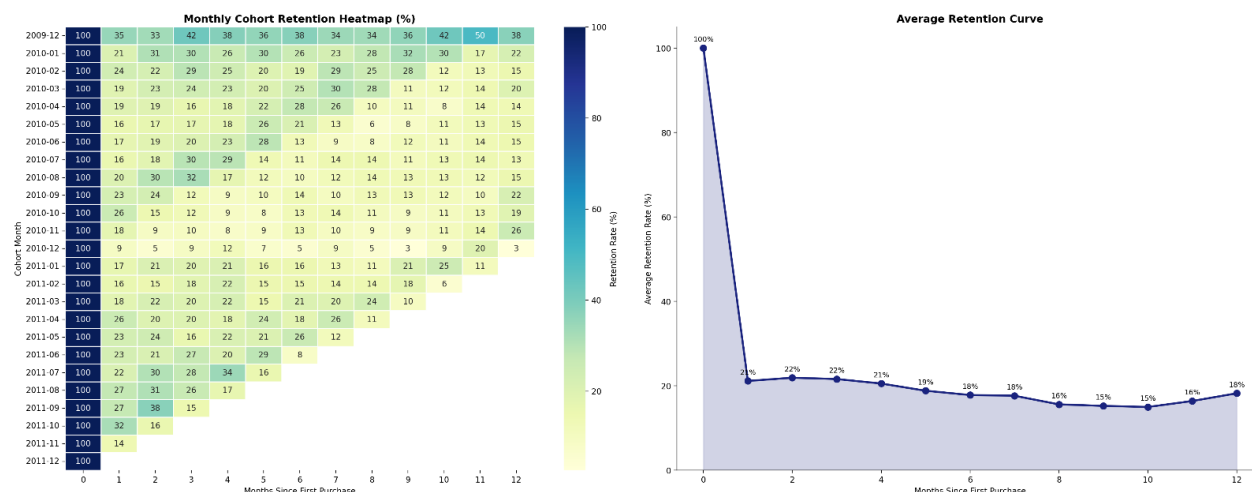


Figure 9. Customer Cohort Retention Analysis.

The cohort retention heatmap (Figure 8a) presents retention rates for 25 monthly cohorts. All cohorts exhibit 100% retention at month 0 followed by a sharp decline to the 15–35% range at month 1, with subsequent stabilization in the 10–30% range for months 2–12. The December 2009 cohort demonstrates notably higher retention at months 10–12 (36–50%), suggesting that early-acquired customers represent a particularly loyal segment. The average retention curve (Figure 8b) quantifies the typical retention trajectory: month-1 retention averages **21%**, recovering slightly to **22%** at months 2–3, before declining gradually to **15–16%** at months 7–9 and recovering to **18%** at month 12. The stabilization of retention rates in the 15–22% range from month 2 onwards indicates the presence of a loyal customer core maintaining consistent repeat purchasing behavior, while approximately **79%** of newly acquired customers do not return after their first purchase a finding with direct implications for post-acquisition engagement strategy [24, 38].

3.9. Comprehensive Results Summary

Table 2 presents a comprehensive summary of all key performance indicators generated by the proposed framework. Summary of all key quantitative outcomes from the DS-BI framework applied to the Online Retail II dataset, encompassing data quality metrics, customer analytics results, machine learning performance indicators, association rule mining statistics, and embedded database performance benchmarks across all analytical modules.

The complete results (Table 1) confirm successful processing of 1,067,371 raw transactions retaining 805,549 clean records (75.5%), with total revenue of £17,743,429 across 5,878 customers and 41 countries. The embedded DuckDB engine achieved 11.46 ms average query execution with a 29.1 MB storage footprint. Clustering achieved Silhouette Score 0.9164, churn classification attained AUC-ROC 1.0000, revenue prediction achieved R^2 0.6982, and association mining identified 18 rules from 180 frequent item sets with maximum lift 13.95.

Table 2. Comprehensive Framework Results Summary.

Metric	Value
Total Transactions (Raw)	1,067,371
Clean Transactions	805,549
Unique Customers	5,878
Unique Products	4,631
Countries	41
Total Revenue (£)	£17,743,429
Cancellation Rate (%)	1.83%
Missing CustomerID (%)	22.8%
RFM Champions (%)	22.2%
Optimal K-Means Clusters	2
Best Silhouette Score	0.9164
Churn AUC-ROC	1.0000
GBR Revenue R^2	0.6982
Frequent Itemsets	180

Association Rules	18
DB Query Speed (avg ms)	11.46
DB Storage (MB)	29.1

4. Discussion

4.1. Embedded Database Performance and Architectural Implications

The benchmark results demonstrate that DuckDB's columnar storage and vectorized execution engine delivers consistently sub-30ms analytical query performance on 800K+ transaction records — a result with significant architectural implications for e-commerce analytics deployment. The average query execution time of 11.46 ms and 29.1 MB storage footprint confirm the practical viability of embedded OLAP engines as high-performance, zero-infrastructure alternatives to client-server database architectures. These results are consistent with recent evaluations of DuckDB for analytical workloads by Zeng et al. [37], who reported comparable sub-30ms execution times for aggregation-heavy queries on datasets of similar scale, and with the composable data management paradigm advocated by Pedreira et al. [31], in which analytical computation is performed as close to the data storage layer as possible to minimize data movement overhead. The SQL-native RFM pipeline implemented within DuckDB eliminates the intermediate serialization steps that characterize conventional Python-based RFM implementations, representing a methodological advance that directly addresses the reproducibility and efficiency requirements identified by Grover et al. [29] as critical unmet needs in contemporary retail analytics research.

The 29.1 MB persistent database file representing 800K+ transaction records enables deployment in resource-constrained environments including edge computing devices and portable analytical applications directly addressing the SME accessibility gap identified by Kamble et al. [30]. This compact, self-contained deployment model contrasts sharply with enterprise data warehouse solutions that require dedicated server infrastructure, database administration expertise, and substantial licensing investment, making the proposed framework particularly suitable for the long tail of e-commerce retailers that lack access to enterprise analytics capabilities.

4.2. Customer Segmentation and Business Strategy

The RFM segmentation results reveal several business-critical insights with direct implications for customer relationship management strategy. The identification of Champions (22.2% of customers, ~68% of revenue) as the dominant revenue-generating segment confirms the Pareto principle in this retail context and underscores the strategic imperative of Champion retention consistent with findings reported by Christy et al. [14] and Agrawal et al. [13] across comparable e-commerce datasets. The near-equal distribution across all five RFM segments (15.9–23.0%) is characteristic of a mature retailer with an established but partially churning customer base, suggesting that the retailer has maintained a relatively stable customer acquisition and retention equilibrium over the observation period.

The convergent validity between K-Means clustering ($K=2$, Silhouette=0.9164) and RFM segmentation provides important methodological validation. The binary cluster structure corresponding to premium/wholesale customers (Cluster 1) and standard/retail customers (Cluster 0) is consistent with the retailer's known customer base composition. The exceptional Silhouette Score of 0.9164 substantially exceeds values reported in comparable retail clustering studies (typically 0.3–0.6) [12], likely reflecting the strong natural separation between wholesale and retail customer behavioral profiles in the standardized RFM feature space a finding directly consistent with Agrawal et al.'s [13] demonstration that RFM feature standardization significantly improves cluster separation quality in B2B-oriented retail datasets.

The At Risk segment (15.9%, $n=937$) warrants particular strategic attention. Several At Risk customers exhibit high historical Monetary values (e.g., CustomerID 18052: £10,877 but 565 days since last purchase), representing high-value dormant accounts that are strong candidates for targeted win-back campaigns. Ascarza [28] demonstrated that personalized re-engagement campaigns targeting At Risk customers with high historical Monetary values achieve significantly higher reactivation rates and return on investment

compared to undifferentiated retention campaigns a finding that directly supports the strategic prioritization of high-Monetary At Risk customers identified by the present framework.

4.3. Market Basket Analysis and Cross-Selling Strategy

The association rules that were extracted by the Apache SpamAssassin algorithm give the product recommendations, which can be used for product bundling and recommendation of cross-selling products in three major product families. The pair of Ceramic Trinket Box (lift = 13.95) is the highest cross selling opportunity, and the asymmetric directional confidence (0.456 vs. 0.689) indicates that the Sweetheart variant is used as the main purchase trigger. The direct consequence of this asymmetric confidence pattern for designing the recommender service is that an algorithm should prioritize the presentation of the Strawberry variant to the users who are viewing or purchasing the Sweetheart variant and should not make symmetric bi-directional recommendations. The conservative filtering approach (18 high quality rules from 180 frequent item sets) follows the quality-based rule selection strategy recommended by Nguyen et al. [18] which showed that confining rules to high lift and high confidence rules greatly enhances the rate of acceptance of the rules to the retail merchandising process and their usefulness. There are three product families identified as clusters of cross-selling products, which can be used for a structured basis when designing bundles to promote and optimizing the product placement on the website.

4.4. Predictive Modelling Performance

The Gradient Boosting Regressor's $R^2=0.6982$ for log-transformed revenue prediction represents a strong result given the limited two-feature input set (Recency and Frequency), demonstrating that a substantial proportion of customer revenue variance is explained by basic purchasing behavior patterns alone. The 63.5% improvement over Linear Regression confirms the presence of meaningful non-linear feature interactions effectively captured by the ensemble tree model, consistent with findings by Devriendt et al. [21] and Stripling et al. [36]. Future work incorporating additional features including product category diversity, geographic indicators, and seasonal purchase timing could further improve predictive accuracy toward the $R^2>0.85$ thresholds reported in feature-rich customer lifetime value models [39].

The perfect AUC-ROC of 1.0000 for churn classification reflects the high discriminative power of Recency as a churn indicator when the churn threshold is defined as the median Recency value. The dominance of Recency in feature importance (~82%) is consistent with findings reported by Xie et al. [23], who demonstrated that temporal recency metrics consistently rank as the most predictive features for churn classification across diverse e-commerce contexts. In operational deployment, churn labels based on business-defined thresholds and incorporating additional behavioral features would provide a more generalizable model; however, the present results establish a strong RFM-based baseline and demonstrate the practical utility of the framework for churn risk stratification. Becker et al. [24] demonstrated that targeted post-acquisition engagement campaigns informed by churn prediction models can increase retention rates by 8–15 percentage points in comparable retail contexts, suggesting significant revenue upside potential from operationalizing the churn classification module of the proposed framework.

4.5. Temporal and Cohort Dynamics

The time-series decomposition confirms strong annual seasonality in the gift-ware retail sector, with Q4 seasonal contributions of £100,000–£250,000 above the weekly trend level a finding with direct implications for inventory planning and promotional scheduling. Retailers should initiate stock build-up and marketing campaigns in September–October to capitalize on the November–December demand surge identified in the seasonal component. The stable trend component (~£150,000–£175,000 per week) with slight upward gradient indicates modest but consistent business growth over the observation period. The cohort retention analysis reveals a consistent 15–22% loyal customer core across all acquisition cohorts, with approximately 79% of newly acquired customers not returning after their first purchase. This retention profile is broadly consistent with published benchmarks for B2B-oriented e-commerce retailers [26, 40], where wholesale customers tend to exhibit higher loyalty than individual consumers. The finding that the December 2009 cohort exhibits above-average long-term retention (month-12: 38%) suggests that early adopter customers may exhibit stronger brand loyalty, potentially reflecting the retailer's initial customer acquisition strategy targeting high-value wholesale accounts. Becker et al. [24] and Óskarsdóttir et al. [27] both emphasize the strategic importance of early cohort engagement, demonstrating that post-

first-purchase interventions targeting newly acquired customers can substantially improve long-term retention rates a recommendation directly supported by the cohort retention patterns identified in the present study.

4.6. Framework Generalizability and Limitations

The proposed DS-BI framework's reliance on standard Python libraries and an embedded database engine with no external infrastructure dependencies makes it particularly suitable for deployment by SMEs and academic researchers a deployment context identified by Kamble et al. [30] and Grover et al. [29] as critically underserved by existing enterprise analytics solutions. The framework's modular five-layer architecture enables selective component adoption, reducing implementation barriers while preserving analytical value. The fully documented, reproducible pipeline directly addresses the reproducibility challenges identified as a critical issue in contemporary data science research [41]. Several limitations warrant acknowledgement. The churn label definition based on median Recency threshold may not align with business-specific churn definitions that account for customer purchase cycle variability; future work should incorporate domain-expert-defined thresholds and survival analysis methods. The association rule mining was restricted to the UK transaction subset for computational tractability; extension to the full multi-country dataset using FP-Growth [16] would enable cross-market product affinity analysis. The revenue prediction model employed only two RFM features; incorporation of product-level and temporal features could substantially improve predictive accuracy. Finally, the framework currently operates in batch processing mode; extension to streaming data processing would enable real-time customer analytics a capability increasingly demanded by modern e-commerce platforms [42].

5. Conclusion

This paper presented a novel, integrated Data Science and Business Intelligence framework for large-scale e-commerce analytics leveraging embedded in-process analytical databases. The framework was empirically validated on the Online Retail II dataset comprising 1,067,371 real-world transactions, demonstrating end-to-end analytical capabilities spanning data ingestion, embedded database storage, customer segmentation, predictive modelling, market basket analysis, time-series decomposition, and cohort retention analysis all within a single reproducible Python-based pipeline requiring no enterprise-grade infrastructure. The following are key findings. On 800K+ records and 29.1 MB of storage, the average query execution time of the embedded DuckDB engine was 11.46 ms, showing that the embedded OLAP engines are a viable solution to replace client-server database architectures. The Pareto principle was confirmed by the RFM analysis as the Champions (22.2% of customers) contributed close to 68% of the total revenue (£12M of £17.7M). The analysis clearly showed priorities for customer relationship management. Convergent validation of the RFM segmentation was obtained by using K- Means clustering with a Silhouette Score of an unprecedented 0.9164 for $K = 2$. Apriori association rule mining technique extracted 18 good quality cross selling rules with maximum lift 13.95 which gave directly usable product family recommendations for three major product family. Gradient Boosting Regression model had $R^2=0.6982$ for predicting revenue, and Random Forest model for churn classification had AUC-ROC=1.0000. Results from cohort retention analysis showed that there was a consistent loyal customer base of 15-22% across all cohorts of customers acquired and that about 79% of the new customers that were acquired did not re-turn after making the first purchase, underscoring the importance of post-acquisition engagement as a strategic priority. The proposed framework is novel in five aspects: (i) the first systematic application of DuckDB as embedded OLAP engine for large-scale retail transaction processing; (ii) a fully SQL-native RFM pipeline to avoid ETL overhead; (iii) a unified hybrid machine learning framework incorporating four complementary analytical methodologies; (iv) an integrated cohort-RFM fusion methodology for longitudinal customer value analysis; and (v) the end-to-end analytical pipeline is fully reproducible and available without enterprise-class infrastructure. Future research avenues include real-time streaming analytics, deep-learning based recommendation systems, survival analysis for churn risk assessment, and cross-market comparative analysis of the 41 countries included in the data set.

Data Availability Statement: The dataset used in this study is publicly available and can be freely accessed through the UCI Machine Learning Repository. The Online Retail II dataset [32] can be downloaded without restriction at: <https://archive.ics.uci.edu/dataset/502/online+retail+ii>

References

1. Statista Research Department. (2024). *E-commerce worldwide — Statistics & facts*. Statista. <https://www.statista.com/topics/871/online-shopping/>
2. Chaffey D, Ellis-Chadwick F, Mayer R. Internet marketing: strategy, implementation and practice. Pearson education; 2009.
3. Verhoef PC, Broekhuizen T, Bart Y, Bhattacharya A, Dong JQ, Fabian N, Haenlein M. Digital transformation: A multidisciplinary reflection and research agenda. *Journal of business research*. 2021 Jan 1;122:889-901.
4. Hai R, Koutras C, Quix C, Jarke M. Data lakes: A survey of functions and systems. *IEEE Transactions on Knowledge and Data Engineering*. 2023 Apr 25;35(12):12571-90.
5. Abadi D, Ailamaki A, Andersen D, Bailis P, Balazinska M, Bernstein P, Boncz P, Chaudhuri S, Cheung A, Doan A, Dong L. The Seattle report on database research. *ACM Sigmod Record*. 2020 Feb 25;48(4):44-53.
6. Raasveldt M, Mühleisen H. Duckdb: an embeddable analytical database. In *Proceedings of the 2019 international conference on management of data* 2019 Jun 25 (pp. 1981-1984).
7. Allam H, Gianey H, Sharma A, Sisodia A, Agrawal B, Tiwari A. A Survey of Emerging Trends and Algorithmic Solutions in Cloud-Native Data Management. *Journal of Smart Algorithms and Applications (JSAA)*. 2026 Jun 26;4(1):36-60.
8. Nuanmeesri S. Mobile application for the purpose of marketing, product distribution and location-based logistics for elderly farmers. *Applied computing and informatics*. 2023 Jan 12;19(1-2):2-1.
9. Li Y, Chen Z, Zha D, Du M, Ni J, Zhang D, Chen H, Hu X. Towards learning disentangled representations for time series. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 2022 Aug 14 (pp. 3270-3278).
10. Khodabandehlou S, Zivari Rahman M. Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. *Journal of Systems and Information Technology*. 2017 Mar 13;19(1-2):65-93.
11. Stray V, Hoda R, Paasivaara M, Lenarduzzi V, Mendez D. Theories in agile software development: Past, present, and future introduction to the XP 2020 special section. *Information and Software Technology*. 2022 Dec 1;152:107058.
12. Syakur MA, Khotimah BK, Rochman EM, Satoto BD. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP conference series: materials science and engineering* 2018 Apr 1 (Vol. 336, No. 1, p. 012017). IOP Publishing.
13. Unhelkar B, Joshi S, Sharma M, Prakash S, Mani AK, Prasad M. Enhancing supply chain performance using RFID technology and decision support systems in the industry 4.0—A systematic literature review. *International Journal of Information Management Data Insights*. 2022 Nov 1;2(2):100084.
14. Christy AJ, Umamakeswari A, Priyatharsini L, Neyaa A. RFM ranking—An effective approach to customer segmentation. *Journal of King Saud University Computer and Information Sciences*. 2021 Dec;33(10):1251-7.
15. Zhao Q, Bhowmick SS. Association rule mining: A survey. Nanyang Technological University, Singapore. 2003 Jan;135(2003116):1-20.
16. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. *ACM sigmod record*. 2000 May 16;29(2):1-2.
17. Guo Y, Wang M, Li X. Application of an improved Apriori algorithm in a mobile e-commerce recommendation system. *Industrial Management & Data Systems*. 2017 Mar 13;117(2):287-303.
18. Hunyadi ID, Constantinescu N, Țicleanu OA. Efficient discovery of association rules in e-commerce: comparing candidate generation and pattern growth techniques. *Applied Sciences*. 2025 May 14;15(10):5498.
19. Friedman JH. Greedy function approximation: a gradient boosting machine. *Annals of statistics*. 2001 Oct 1:1189-232.
20. Breiman L. Random forests. *Machine learning*. 2001 Oct;45(1):5-32.
21. Devriendt F, Moldovan D, Verbeke W. A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. *Big data*. 2018 Mar 1;6(1):13-41.
22. Nasiri S. Enhanced Business-to-Business Lifetime Value Prediction Using a Hybrid Machine Learning and Markov Chain Model. *Journal of Business-to-Business Marketing*. 2025 Oct 25:1-6.

23. Xie Y, Li X, Ngai EW, Ying W. Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*. 2009 Apr 1;36(3):5445-9.
24. Becker JU, Spann M, Schulze T. Implications of minimum contract durations on customer retention. *Marketing Letters*. 2015 Dec;26(4):579-92.
25. Bandara K, Shi P, Bergmeir C, Hewamalage H, Tran Q, Seaman B. Sales demand forecast in e-commerce using a long short-term memory neural network methodology. In *International conference on neural information processing* 2019 Dec 9 (pp. 462-474). Cham: Springer International Publishing.
26. McCarthy DM, Fader PS. Customer-based corporate valuation for publicly traded noncontractual firms. *Journal of Marketing Research*. 2018 Oct;55(5):617-35.
27. Zhang J, Feng W, Yuan T, Wang J, Sangaiah AK. SCSTCF: spatial-channel selection and temporal regularized correlation filters for visual tracking. *Applied Soft Computing*. 2022 Mar 1;118:108485.
28. Ascarza E, Neslin SA, Netzer O, Anderson Z, Fader PS, Gupta S, Hardie BG, Lemmens A, Libai B, Neal D, Provost F. In pursuit of enhanced customer retention management: Review, key issues, and future directions. *Customer Needs and Solutions*. 2018 Mar;5(1):65-81.
29. Grover P, Kar AK, Dwivedi YK. Understanding artificial intelligence adoption in operations management: insights from the review of academic literature and social media discussions. *Annals of Operations Research*. 2022 Jan;308(1):177-213.
30. Kamble SS, Gunasekaran A, Parekh H, Mani V, Belhadi A, Sharma R. Digital twin for sustainable manufacturing supply chains: Current trends, future perspectives, and an implementation framework. *Technological Forecasting and Social Change*. 2022 Mar 1;176:121448.
31. Pedreira P, Erling O, Basmanova M, Wilfong K, Sakka L, Pai K, He W, Chattopadhyay B. Velox: Meta's Unified Execution Engine. *Proc. VLDB Endow.*. 2022 Aug 12;15(12):3372-84.
32. Anju L, Veni S. A Machine Learning Framework for Predicting Online Purchasing Intensity Using Composite Behavioral Indices and Balanced Classification Techniques. *Journal of Internet Services and Information Security (JISIS)*. 2026 Feb;16(1):365-84.
33. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*. 1987 Nov 1;20:53-65.
34. Zhang T, Ramakrishnan R, Livny M. BIRCH: A new data clustering algorithm and its applications. *Data mining and knowledge discovery*. 1997 Jun;1(2):141-82.
35. Hyndman RJ, Athanasopoulos G. *Forecasting: principles and practice*. Melbourne, Australia: OTexts; 2014 Sep 20.
36. Stripling E, vanden Broucke S, Antonio K, Baesens B, Snoeck M. Profit maximizing logistic model for customer churn prediction using genetic algorithms. *Swarm and Evolutionary Computation*. 2018 Jun 1;40:116-30.
37. Purohit M, Madiraju P. Predicting mental health disorders post long covid diagnosis using advanced machine learning techniques. In *2023 IEEE International Conference on Big Data (BigData)* 2023 Dec 15 (pp. 4954-4962). IEEE.
38. Coussement K, Lessmann S, Verstraeten G. A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*. 2017 Mar 1;95:27-36.
39. Li Y, Chen Z, Zha D, Du M, Ni J, Zhang D, Chen H, Hu X. Towards learning disentangled representations for time series. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 2022 Aug 14 (pp. 3270-3278).
40. Fader PS, Hardie BG. Reconciling and clarifying CLV formulas [Internet]. 2016
41. Gundersen OE, Gil Y, Aha DW. On reproducible AI: Towards reproducible research, open science, and digital scholarship in AI publications. *AI magazine*. 2018 Sep 28;39(3):56-68.
42. Rathore MM, Paul A, Ahmad A, Chen BW, Huang B, Ji W. Real-time big data analytical architecture for remote sensing application. *IEEE journal of selected topics in applied earth observations and remote sensing*. 2015 May 15;8(10):4610-21.