Quantitative Finance and Economics

QFE, 9(3): 682–712.
DOI: 10.3934/QFE.2025024
Received: 07 April 2025
Revised: 21 July 2025
Accepted: 20 August 2025
Published: 15 September 2025

https://www.aimspress.com/journal/QFE

Research article

Big data and consumer behavior: A macroeconomic perspective through supermarket analytics

Tasos Stylianou[1],* and Aikaterina Pantelidou[2]

[1] Department of Economics, University of Macedonia, Thessaloniki, Greece
[2] School of Computing, Mediterranean College, Thessaloniki, Greece

* **Correspondence:** Email: tasosstylianou@gmail.com, tstyl@uom.edu.gr.

Abstract: This study explores how big data analytics can be used on supermarket transaction data to reveal patterns in consumer behavior with broader macroeconomic implications. Using a comprehensive dataset from a multinational retail chain, we employed advanced analytical methods—including machine learning algorithms, time series forecasting (ARIMA), clustering, and recommendation systems—to model purchasing behavior and segment customers. The analysis revealed distinct consumer profiles, habitual spending patterns, and time-sensitive trends that inform both retail decision-making and economic interpretation. The results demonstrated that consumer transaction data can serve not only to improve operational efficiency and personalized marketing in the retail sector but also to provide real-time indicators of economic sentiment and household financial health. This dual contribution highlights the potential of retail big data as a tool for both business strategy and macroeconomic policy analysis. The study also outlines implementation considerations and ethical challenges, offering a foundation for future research in data-driven retail and economic analytics.

Keywords: big data analytics; consumer behavior; supermarket analytics; machine learning; macroeconomic analysis

JEL Codes: C45, C81, D12, L81

1. Introduction

The exponential growth of data in today's digital economy has positioned big data as a cornerstone for strategic decision-making across various sectors. Among these, the retail industry has particularly benefited from the ability to harness consumer data to inform marketing strategies, operational logistics, and customer engagement initiatives (McAfee et al., 2012; Chen et al., 2012). Big data encompasses vast, high-velocity, and high-variety datasets generated through consumer interactions, such as in-store purchases, online transactions, loyalty programs, and digital footprints, which—when properly analyzed—can reveal nuanced patterns of consumer behavior (Gandomi and Haider, 2015).

Big data is characterized by the five Vs: volume, velocity, variety, veracity, and value (Zikopoulos and Eaton, 2011), although contemporary frameworks often extend this to include additional dimensions such as *variability*, *visualization*, and *temporality*. These attributes collectively define the challenges and opportunities in collecting, processing, and interpreting massive datasets. The temporal aspect of big data—how information changes over time—is particularly important in retail analytics, where consumer behavior is dynamic and influenced by seasonal patterns, economic shifts, and marketing interventions. Temporality interacts closely with *veracity* (the reliability of data across periods) and *variety* (as new data sources and formats emerge with time), adding complexity to predictive modeling and longitudinal analysis.

In the retail sector, and supermarkets in particular, these datasets encompass millions of timestamped transactions enriched with geographic, behavioral, and demographic variables. Analyzing this high-dimensional, time-sensitive data requires advanced analytical techniques. Machine learning methods such as clustering, association rule mining, and time series forecasting are increasingly deployed to translate raw inputs into meaningful consumer insights (Fayyad et al., 1996; Hastie et al., 2009).

The analysis of consumer behavior using big data has emerged as a key area of interest in marketing science and behavioral economics (Deaton and Muellbauer, 1980; Germann et al., 2014). Consumer behavior is influenced by multiple interacting variables, including product preferences, price sensitivity, promotional exposure, and broader macroeconomic factors. Studies show that integrating these behavioral patterns into data models can significantly enhance forecasting accuracy and customer satisfaction (Corrigan et al., 2014; Rooderkerk et al., 2013). From a macroeconomic perspective, shifts in consumption patterns can also act as indicators of economic health, enabling real-time policy insights (Einav and Levin, 2014).

Supermarket analytics is a practical and data-rich field in which consumer behavior is captured at scale. Leading global retailers, such as Walmart, process billions of transactions annually, enabling them to dynamically adjust supply chains, product assortments, and pricing strategies (Kambatla et al., 2014; Kamel, 2023). This study builds on these developments by analyzing consumer transaction data from a multinational supermarket chain. We apply machine learning algorithms and statistical models to derive a granular understanding of purchasing behavior and assess its implications for macroeconomic trends.

While prior studies have explored the micro-level benefits of big data in retail, such as personalization and promotion optimization (Giri et al., 2019; Aktas and Meng, 2017), there remains a research gap in connecting these insights with macroeconomic indicators. This study contributes to filling this gap by examining how variations in product category preferences, purchasing frequencies, and transaction volumes reflect broader economic conditions such as inflationary pressure or consumer confidence (Aguiar and Hurst, 2005; Deaton and Muellbauer, 1980).

This research is grounded in the growing body of interdisciplinary literature at the intersection of retail analytics, data science, and economic policy. It proposes a structured framework combining machine learning, clustering, and time series analysis to identify patterns that inform both firm-level strategy and economic planning. The results have practical relevance for retail managers, marketers, economists, and policymakers seeking to understand and forecast consumer trends in a data-driven economy.

1.1. Research question and objectives

Despite extensive literature on big data and retail analytics, most existing studies focus on firm-level benefits such as customer targeting or supply chain optimization (Giri et al., 2019; Aktas and Meng, 2017). There remains a research gap in assessing how consumer transaction data can be used to uncover macroeconomic signals. This study addresses the following research questions:

- **RQ1:** How can big data analytics techniques be employed to identify patterns and trends in supermarket consumer behavior?
- **RQ2:** What are the macroeconomic implications of these patterns, particularly regarding consumer sentiment and financial health?
- **RQ3:** How can customer segmentation and recommendation models be integrated into retail strategies to enhance operational efficiency and customer experience?

To address these questions, this study uses real-world transaction data from a retail supermarket chain and applies various machine learning and statistical techniques—including clustering, association rule mining, and time series forecasting.

1.2. Novelty and contribution of the study

This paper contributes to the literature by offering both methodological and conceptual novelties. First, it integrates unsupervised and supervised learning techniques within a unified framework to model consumer behavior. Second, it links micro-level purchasing patterns to macro-level economic interpretations, positioning consumer data as a tool for economic insight. Finally, the study introduces a scalable, data-driven model for retailers to improve their personalization strategies through customer segmentation and predictive modeling.

Most importantly, while previous work has explored either consumer analytics or macroeconomic forecasting in isolation, this paper builds a bridge between the two domains using big data. The macroeconomic framing of consumer behavior through supermarket analytics is a unique contribution that enhances both academic understanding and applied decision-making.

The analysis of big data for understanding consumer behavior holds significant implications for both retailers and policymakers. For retailers, the ability to predict consumer behavior and adjust strategies accordingly can result in increased sales, improved customer satisfaction, and a strengthened competitive advantage. Meanwhile, policymakers can utilize insights into consumer spending patterns to inform economic policies aimed at driving growth, managing inflation, and ensuring economic stability.

Despite the growing literature on big data in the retail industry, the majority of studies remain focused on firm-level outcomes, such as personalization, marketing optimization, and operational efficiencies (Kamel, 2023; Germann et al., 2014). These microeconomic applications, while valuable, overlook the broader question of how consumer behavior—when aggregated and analyzed at scale—can serve as a signal for macroeconomic conditions such as consumer confidence, inflation expectations, and household financial health.

This paper addresses this research gap by analyzing consumer purchasing patterns using supermarket data and interpreting these trends within a macroeconomic framework. In doing so, it extends the scope of big data analytics from business intelligence to economic insight. This dual focus—linking micro-level retail data to macroeconomic indicators—is the core novelty of our study.

By integrating advanced machine learning techniques, such as clustering, recommendation systems, and time series forecasting, with an economic interpretive lens, this research makes a methodological and conceptual contribution to the interdisciplinary field of data-driven economic analysis. The study also demonstrates how such models can be implemented at scale to provide real-time decision support for businesses and public institutions alike.

1.3. Structure of the paper

The rest of this paper is structured as follows: Section 1 introduces the topic, outlines the research questions and objectives, discusses the study's novelty, and sets the conceptual background. Section 2 provides a comprehensive literature review, synthesizing relevant studies on big data analytics, retail analytics, and customer behavior modeling, while highlighting theoretical and practical gaps. Section 3 presents the methodology, including data description, preprocessing techniques, exploratory data analysis, and machine learning algorithms used (Apriori, collaborative filtering, K-means clustering, and ARIMA). Section 4 discusses the results, including visualizations and insights on shopping behavior patterns, segmentation, and recommendation model performance. Section 5 offers a discussion, comparing the findings with existing literature and elaborating on their implications for retailers and policymakers. Section 6 concludes the paper, summarizing the key contributions and offering recommendations for retail practice and economic policy. It also outlines future research directions and addresses limitations.

2. Literature review

2.1. Big data and analytics: conceptual foundations

Big data (BD) refers to datasets that are high in volume, velocity, and variety, characteristics that render traditional data processing techniques inadequate. The extension of the 3Vs model into 5Vs—adding veracity (accuracy and reliability) and value (potential utility of data)—has become standard in the literature to underscore the challenges and promises of BD (Gandomi and Haider, 2015; Watson, 2014). The effective handling of BD requires advanced computational infrastructure, scalable storage, and robust analytical methodologies. These include distributed computing frameworks such as Hadoop and Spark, as well as advanced machine learning algorithms.

BD is generally categorized into structured (e.g., transactional data), semi-structured (e.g., JSON, XML), and unstructured data (e.g., videos, social media content), all of which are pertinent in the retail context. The complexity of managing and integrating these diverse data formats necessitates a systematic analytical approach.

2.2. Evolution of big data analytics (BDA) in retail

In the retail sector, BDA has evolved from simple transactional analysis to sophisticated systems capable of predictive and prescriptive insights. Early BDA frameworks focused on descriptive analytics (e.g., sales summaries), whereas modern systems increasingly integrate predictive (e.g., forecasting demand) and prescriptive analytics (e.g., determining optimal pricing strategies) (Chen et al., 2012; McAfee et al., 2012).

Retailers such as Amazon, Walmart, and Tesco have leveraged BDA to improve operational efficiency, customer experience, and strategic decision-making. These improvements include demand forecasting, inventory optimization, dynamic pricing, and personalized marketing. For example, Walmart's data platform processes over 2.5 petabytes of data hourly, enabling real-time business intelligence (Kamel, 2023). BDA also supports assortment planning by analyzing product relationships and identifying substitution effects across categories (Rooderkerk et al., 2013).

Furthermore, BDA facilitates advanced promotion analysis and price optimization. Retailers analyze historical sales data, competitor pricing, and consumer responses to identify optimal pricing points and promotional strategies. These models integrate econometric principles with machine learning algorithms, enhancing both accuracy and scalability.

2.3. Contemporary models in big data analytics for customer analytics

The complexity of modern consumer behavior necessitates advanced analytical frameworks that go beyond traditional models. Recent literature introduces several contemporary BDA models for customer analytics:

- CRISP-DM (cross-industry standard process for data mining). This six-phase model (business understanding, data understanding, data preparation, modeling, evaluation, and deployment) remains one of the most practical and widely adopted BDA frameworks. Its iterative nature aligns well with the continuous nature of retail data collection and decision-making (Wirth and Hipp, 2000). CRISP-DM facilitates systematic exploration of customer data, allowing teams to align analytics with business goals effectively.
- Enhanced customer lifetime value (CLV) modeling. Modern CLV models incorporate machine learning techniques like gradient boosting, XGBoost, and neural networks to improve the accuracy of customer value estimation. Unlike traditional RFM-based models, ML-enhanced CLV predictions allow for better personalization and segmentation strategies (Smaili and Hachimi, 2023; Kumar and Petersen, 2005). Such approaches also support dynamic updating of CLV as new transaction data becomes available.
- Deep clustering for segmentation. Customer segmentation has evolved from K-means to more advanced deep clustering methods. Autoencoders, variational autoencoders (VAE), and deep embedded clustering (DEC) are increasingly used to capture nonlinear patterns in high-dimensional

customer data (Xiao et al., 2023). These models allow for flexible and dynamic grouping of customers, particularly useful in fast-changing environments such as online retail.

- Neural collaborative filtering in recommender systems. Recommender systems have transitioned from memory-based collaborative filtering to neural collaborative filtering (NCF), which combines matrix factorization with deep learning to model complex user–item interactions (He et al., 2017). Hybrid recommendation engines that integrate content, collaborative, and contextual data offer the most accurate personalization. NCF models outperform traditional systems in accuracy, cold-start problem mitigation, and scalability.
- Customer journey analytics (CJA). CJA uses sequential models like hidden Markov models (HMMs) and long short-term memory (LSTM) networks to understand customer paths across channels and touchpoints. This approach is vital in omnichannel retail, where customer interactions span websites, mobile apps, and physical stores (Santos and Gonçalves, 2024). CJA enables real-time behavioral tracking and offers predictive capabilities for customer churn and conversion probabilities.
- Marketing attribution modeling. Marketing attribution models are increasingly integrated with machine learning to assign value to different touchpoints along the customer journey. Data-driven attribution models, such as Shapley value and Markov chain attribution, now replace simplistic first-touch or last-touch models. These methods provide granular insights into the effectiveness of individual campaigns and platforms (Alhamed and Rahman, 2023).

An increasingly vital dimension of customer analytics is temporality—how customer behavior evolves over time across channels and campaigns. Models such as LSTM-based customer journey analytics (CJA) are well-suited to capturing these temporal dependencies, enabling dynamic personalization and churn prediction in real time (Santos and Gonçalves, 2024).

2.4. Implementation challenges and emerging opportunities

Despite its advantages, BDA implementation in retail faces significant challenges. These include data silos, privacy concerns under GDPR and CCPA, talent shortages, and high infrastructure costs. Ensuring data quality across multiple sources is another significant challenge. Integrating real-time and batch data, standardizing taxonomies, and handling missing or inconsistent values require robust data governance frameworks.

However, the emergence of cloud-based platforms (e.g., AWS, Azure, Google Cloud), real-time data lakes, and autoML tools has begun to democratize access to BDA capabilities. Moreover, the adoption of edge computing and federated learning offers promise for processing customer data closer to the source, reducing latency, and enhancing data privacy.

Ethical considerations around algorithmic bias, surveillance, and data ownership are increasingly discussed in academic and policy-making circles. Responsible AI and explainable machine learning are emerging as essential components of any modern BDA framework (Zwitter, 2014; Gandomi and Haider, 2015). Organizations must implement fairness-aware algorithms and maintain transparency in automated decision-making processes.

Recent literature reinforces the shift toward integrated and AI-powered analytics frameworks. Stylianou and Milidis (2024) emphasized the role of socioeconomic and behavioral data fusion in enhancing retail decision-making. Santos and Gonçalves (2024) demonstrated how CJA can be integrated into real-time decision systems. He et al. (2017) validated the superior performance of NCF in dynamic recommendation scenarios.

Additionally, research by Ascarza (2018) and Wang et al. (2025) explored the integration of uplift modeling and causal inference in targeted marketing. These techniques help identify customers whose behavior is most likely to be influenced by specific interventions, improving campaign ROI. Theoretical frameworks such as the Unified Theory of Acceptance and Use of Technology (UTAUT) and Technology-Organization-Environment (TOE) also offer valuable lenses for understanding the adoption and impact of BDA technologies.

Recent contributions further advance this perspective. For instance, Nomura et al. (2025) investigated the application of deep reinforcement learning in personalized pricing, showing how adaptive models can increase revenue and customer satisfaction. Liu et al. (2024) analyzed real-time clickstream data to predict churn behavior using attention-based neural networks, confirming improved accuracy over traditional recurrent models. Moreover, Plebani et al. (2023) integrated blockchain with BDA to enhance data traceability and trust in customer profiling, especially in privacy-sensitive environments.

Emerging work also explores the intersection of BDA with sustainability and corporate social responsibility. Retailers are increasingly using BDA to monitor supply chain emissions, track ethical sourcing, and promote green consumer behavior, signaling a broader strategic alignment between analytics and ESG (environmental, social, and governance) goals. Trebbin and Geburt (2024) presented a hybrid framework combining BDA and lifecycle assessment models to support eco-labeling and carbon footprint transparency in retail supply chains.

Collectively, these studies highlight that contemporary customer analytics is no longer confined to segmentation and recommendation but spans adaptive pricing, trust assurance, and sustainability. Future research should expand upon these multi-dimensional applications and explore the governance models required to manage their complexity and ethical implications.

The literature unequivocally demonstrates that the field of big data analytics has progressed from basic exploratory tools to complex, predictive, and prescriptive frameworks. The integration of AI-driven models such as deep clustering, neural collaborative filtering, and journey analytics provides unparalleled insights into consumer behavior. These tools support hyper-personalization, operational optimization, and strategic planning.

However, effective implementation hinges on overcoming data integration and privacy challenges, adopting ethical frameworks, and ensuring scalability. As BDA technologies continue to evolve, future research should investigate hybrid modeling approaches, the role of human-AI collaboration, and the integration of BDA with broader socio-economic indicators.

In sum, contemporary models in BDA for customer analytics not only redefine how retailers understand and engage with customers but also serve as critical enablers of macroeconomic monitoring, strategic agility, and sustainable growth.

2.4.1. Time series components and theoretical foundations

In retail consumer analytics, time series modeling provides a powerful framework for capturing the temporal dynamics of purchasing behavior. This study adopts a non-seasonal ARIMA(2,1,1) model, which integrates autoregressive lags, differencing, and a moving average term to represent short-term behavioral dependencies. Each component of this model is theoretically grounded in both economic theory and behavioral retail analytics.

The autoregressive (AR) terms, particularly at lags t–1 and t–2, capture habitual purchasing patterns and short-term memory effects. Consumers often exhibit stable and repetitive behavior in their shopping routines, driven by consistent needs, brand loyalty, and household replenishment cycles. According to Einav and Levin (2014), such inertia in consumption is prevalent in household panel data, where recent purchases are strong predictors of near-future demand. Germann et al. (2014) further support this view, demonstrating that past purchase behavior informs future decision-making more reliably than promotional triggers or external stimuli in grocery contexts. The presence of two AR lags in the model reflects both immediate behavioral continuity and slightly longer-term restocking habits, such as biweekly purchasing cycles.

The differencing operator (I) in the ARIMA model is used to transform the original series into a stationary process by removing linear trends. In consumer retail contexts, differencing accounts for evolving consumption baselines, such as gradual increases in demand due to product popularity or economic changes. It ensures that the underlying model captures relative fluctuations rather than absolute levels, which is crucial when modeling time-sensitive, nonlinear shopping behavior.

The moving average (MA) component addresses short-term volatility in demand by capturing the impact of recent, unanticipated shocks. These may include one-time promotions, unexpected stockouts, or situational changes such as weather anomalies or calendar effects. Fayyad et al. (1996) and Chen et al. (2012) emphasized the importance of incorporating error correction mechanisms in models to reflect consumers' adaptive behavior in response to such disturbances. The MA term thus represents consumers' short-term compensatory adjustments, e.g., purchasing more in the current period due to a missed transaction in the previous one.

Taken together, the ARIMA(2,1,1) structure offers a concise yet robust model of non-seasonal, short-run purchasing dynamics. It captures both behavioral momentum and short-term corrections without assuming strict periodicity. This approach is particularly suitable for modern retail datasets where consumer behavior is shaped by rapidly changing preferences, irregular promotional activity, and limited predictability in daily or weekly rhythms. By grounding each component of the model in both theoretical and empirical literature, this study provides a transparent, behaviorally informed foundation for forecasting consumer demand in the supermarket sector. The model's ability to reflect real-time responsiveness and habitual persistence makes it a practical tool for inventory management, pricing strategy, and operational planning.

2.5. Challenges and future directions

The retail sector stands to significantly benefit from the potential of big data (BD) and business data analytics (BDA). However, retailers encounter various challenges in effectively leveraging these technologies. Ensuring data privacy and security is a major concern, as retailers are tasked with safeguarding and ethically using customer data. Adhering to regulations such as GDPR requires robust data governance practices (Kumar and Petersen, 2005). Furthermore, the complexities of integrating data from diverse sources and ensuring its quality necessitate sophisticated data management systems and skilled personnel (Otto et al., 2020).

Looking ahead, the future of BD and BDA in retail holds immense promise. Advancements in artificial intelligence and machine learning will enhance the capabilities of BDA, enabling more precise predictions and deeper insights into customer behavior. The adoption of Internet of Things (IoT) technologies will produce even more data, offering retailers real-time insights into customer interactions and preferences. Additionally, the emergence of advanced analytics platforms and tools will make BDA more accessible to retailers of all sizes, democratizing the advantages of data-driven decision-making. In conclusion, BD and BDA, particularly business data analytics, hold the potential to bring about transformative changes in the retail industry. Through harnessing the power of data, retailers can gain a deeper understanding of their customers, optimize their operations, and drive business growth. The integration of these technologies into retail strategies is essential for maintaining competitiveness in a rapidly evolving market. As technology continues to progress, the role of BD and BDA in retail will become even more critical, offering new opportunities for innovation and success.

3. Methodology and data

The core objective is to analyze transactional data to establish customer purchase behavior profiles. This will ultimately inform and enhance decision-making processes across various aspects of the retail business, including inventory management, supply chain optimization, and in-store product placement strategies. Furthermore, the framework incorporates the development of a recommendation system and customer segmentation based on purchasing habits. This comprehensive approach aims to empower the supermarket chain with actionable insights to drive targeted marketing strategies and informed business decisions.

3.1. Data collection and description

The dataset utilized in this analysis was sourced from Kaggle, a prominent data science competition platform and online community for data scientists and machine learning practitioners (Kaggle, 2023). Kaggle provides a repository for users to find and publish datasets, collaborate with other data scientists, and participate in competitions to address data science challenges. The specific dataset, "ECommerce_consumer behavior", contains online transactional data for a supermarket chain, Hunter, which operates across 10 countries. The dataset comprises 2,019,501 observations and 12 variables, encompassing qualitative and quantitative data. Key variables include order ID, user ID, and products ordered. Each order ID is repeated multiple times within the dataset, as each row represents a single product purchased within an order. This format enables us to determine the quantity of each item bought per order, with each unique order spanning multiple rows corresponding to the number of different products purchased.

3.2. Data preprocessing: foundation for reliable analysis

Data preprocessing serves as the cornerstone of any data analysis endeavor. It encompasses a series of techniques designed to clean, transform, and prepare data for subsequent analysis, ensuring its quality and reliability. Real-world datasets often harbor imperfections that can significantly impact the validity of the analysis. These imperfections can be broadly categorized into three main challenges (Famili et al., 1997).

The data excess problem arises when there are too many noisy or distorted data points and its causes may include inaccurate data entry, measurement errors, or transmission issues. Additionally, for high-dimensional datasets (those with many features), dimensionality reduction techniques may be necessary to facilitate effective data management. Data incompleteness relates to missing or insufficient data points within the dataset. To address these gaps, appropriate imputation methods are essential, as missing variables can significantly impede the analysis process. Data inconsistency occurs when information is obtained from multiple sources or when it is presented in incompatible formats. Inconsistent data formats create significant challenges during preprocessing, making data standardization procedures necessary to ensure uniformity.

Addressing these challenges is paramount to guaranteeing the accuracy and integrity of the data analysis. A variety of techniques are employed during preprocessing, including data cleaning to remove errors and inconsistencies, handling missing values through imputation methods, and standardizing data formats to ensure compatibility. By meticulously attending to these crucial steps, we aim to transform raw data into a high-quality resource that yields reliable and insightful results.

3.3. Exploratory data analysis

Exploratory data analysis (EDA) constitutes an indispensable initial step in the data science workflow, empowering researchers to delve into the intricacies of datasets (Unwin, 2010). This process emphasizes visual and statistical exploration to unearth latent patterns, trends, and relationships within the data, often conducted without preconceived notions. EDA leverages a diverse toolkit of data visualization techniques, such as histograms, scatter plots, and heat maps, to create intuitive representations of the data. These visualizations facilitate the identification of potential correlations and anomalies that might otherwise remain obscured (Li Vigni et al., 2013).

In the context of this supermarket case study, EDA was employed to gain a comprehensive understanding of the customer transaction data. Through the generation of various visualizations, we explored patterns, trends, and relationships among key variables. These variables encompassed purchase frequency by day of the week, the popularity of distinct product categories, and potential correlations between departments frequented by customers. The insights gleaned from this initial exploratory phase laid the groundwork for the subsequent application of more sophisticated machine learning algorithms. This foundational exploration served to accurately define customer preferences and buying habits, ultimately informing the development of targeted strategies for the supermarket chain.

3.4. Machine learning algorithms

Machine learning, a subset of artificial intelligence, involves developing algorithms that enable computers to learn from data and make predictions or decisions without explicit programming (Bata, 2020). This study employs both supervised and unsupervised machine learning algorithms to analyze customer data.

The Apriori algorithm is used for frequent itemset mining and association rule mining in transactional databases. It helps uncover interesting relationships and correlations among products, which can inform decisions on catalog design, store layout, cross-marketing, and promotional strategies (Chang and Liu, 2011; Raeder and Chawla, 2011). The algorithm identifies frequent items and extends them into larger itemsets, provided they appear frequently enough. Key measures include support, confidence, and lift, which help evaluate the strength and reliability of association rules (Aggarwal et al., 2014).

Recommender systems apply statistical and knowledge discovery techniques to provide customized recommendations based on past data (Pravani et al., 2020). This study aims to build a recommendation system for Hunter's supermarket to suggest products based on items added to the cart. Various algorithms, including association rule-based, randomly chosen items, popular items, and collaborative filtering (user-based and item-based), are evaluated based on performance metrics such as precision and recall.

The K-means clustering algorithm is an unsupervised clustering algorithm that identifies homogeneous subgroups within a dataset (Sinaga and Yang, 2020). In this study, K-means clustering is used to segment customers based on their purchase history, allowing the supermarket to implement targeted marketing strategies for each cluster, thereby enhancing customer engagement and increasing revenue.

3.5. Mathematical model for the study

To model consumer behavior in the supermarket context, we can use a combination of the following components:

3.5.1. Transaction data representation

$$T = \{t_1, t_2, \dots, t_n\} \tag{1}$$

where T represents the set of all transactions, and each t_i is a transaction consisting of multiple items purchased by a customer.

3.5.2. Customer segmentation (K-means clustering)

The K-means algorithm is used to segment customers based on their purchasing behavior. The objective function minimized by K-means is:

$$argmin_C \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{2}$$

where C_i is the set of customers in cluster i, μ_i is the centroid of cluster i, and $||x - \mu_i||^2$ is the squared Euclidean distance between a customer x and the cluster centroid μ_i.

3.5.3. Frequent itemset mining (Apriori algorithm)

The Apriori algorithm identifies frequent itemsets, where the support for an itemset I is defined as:

$$Support(I) = \frac{Number\ of\ transactions\ containing\ I}{Total\ Number\ of\ transactions} \quad (3)$$

Association rules are then derived from these frequent itemsets with confidence and lift metrics:

$$Confidence(A \rightarrow B) = \frac{Support(A \cup B)}{Support(A)}$$

$$Lift(A \rightarrow B) = \frac{Confidence(A \rightarrow B)}{Support(B)} \quad (4)$$

3.5.4. Predictive modeling (time series analysis)

Time series analysis is used to predict future sales trends based on historical purchasing patterns. In this study, we employed an autoregressive integrated moving average (ARIMA) model to analyze daily transaction data. The general ARIMA model is expressed as:

$$Y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \theta_1 \varepsilon_{t-1} + \varepsilon_t \quad (5)$$

where Y_t is the observed value (e.g., sales) at time t, φ_1 and φ_2 are the parameters of the autoregressive (AR) terms at lags1 and 2, θ_1 is the parameter of the moving average (MA) term at lag 1, ε_t is the error term, and c is the constant or intercept term.

The ARIMA(2,1,1) specification used in this study was selected based on AIC/BIC model selection criteria and evidence of significant autocorrelation in the data. The model captures short-run momentum, behavioral inertia, and corrective feedback mechanisms in consumer purchasing behavior.

Model insights and key variables:

- **Lagged purchases (Y_{t-1})** had the strongest predictive effect, confirming that consumer behavior is largely habitual and strongly influenced by recent purchase activity. This supports theories of routine-based consumption and behavioral reinforcement (Einav and Levin, 2014).
- **Second-lag purchases (Y_{t-2})** also showed significant influence, capturing longer short-term memory, such as biweekly restocking cycles or spillover from previous promotional purchases (Wang et al., 2025).
- **Short-term shocks (ε_{t-1}),** while significant, had a comparatively smaller effect. These reflect corrective behaviors following disruptions like promotions, stockouts, or unexpected changes in household needs (Hastie et al., 2009; Nomura et al., 2025).

This mathematical framework integrates predictive modeling with consumer behavior theory to provide a robust understanding of demand evolution in supermarket settings. Unlike seasonal models, ARIMA(2,1,1) is more appropriate in contexts where daily or weekly patterns are irregular or not systematically repeated.

The ARIMA model provides actionable insights for retail strategy, particularly in inventory management, short-term marketing response planning, and real-time demand monitoring. Forecasts derived from this model can help managers make proactive decisions based on behavioral momentum rather than reacting to sales volatility.

3.5.5. Research hypotheses

To deepen the analysis of consumer purchasing behavior through time series modeling, this study formulates explicit research hypotheses based on the components of the ARIMA(2,1,1) model applied in Section 3.5.4. The ARIMA model captures non-seasonal temporal dependencies by combining autoregressive, differencing, and moving average components. The hypotheses reflect theoretically grounded behavioral expectations regarding persistence in consumer habits and their sensitivity to short-term irregularities.

H1: Lagged consumer purchases (Y_{t-1} , Y_{t-2}) significantly influence current consumer demand patterns.

This hypothesis is based on the notion that purchasing behavior is habit-driven and path-dependent. Consumers often repeat similar purchasing behavior over consecutive weeks due to routine household needs, brand loyalty, and replenishment patterns. Research in retail behavior confirms that previous purchase activity is a strong predictor of near-future transactions (Einav and Levin, 2014; Germann et al., 2014). Recent empirical evidence further supports this autoregressive tendency in grocery settings (Wang et al., 2025).

H2: Previous period's forecast error (ε_{t-1}) significantly affects current consumer demand.

This hypothesis posits that consumers partially compensate for short-term deviations or shocks. For example, if a household makes an unexpectedly small purchase one week, it may make up for it in the next. This behavior is captured by the moving average component of the ARIMA model. Theoretical backing comes from learning-based models and dynamic behavior frameworks, suggesting that consumers adjust to deviations from normative patterns (Fayyad et al., 1996; Hastie et al., 2009; Nomura et al., 2025).

By validating these hypotheses within the ARIMA framework, the study aims to interpret short-term behavioral inertia and responsiveness to recent deviations in the context of supermarket shopping. The goal is to increase predictive reliability while also offering a clearer theoretical explanation for demand evolution in routine retail environments.

3.6. Model implementation

Apriori algorithm implementation: The Apriori algorithm was utilized to identify frequent itemsets and extract association rules from the transactional data. Key metrics—support, confidence, and lift—were calculated to evaluate the strength and significance of these associations, which informed decisions regarding product placement and promotional strategies.

Recommended system implementation: Several recommendation algorithms were assessed to determine the most effective method for suggesting products to customers. Special attention was given to user-based and item-based collaborative filtering techniques, which leverage similarities in customer preferences and purchase histories. These approaches showed superior performance in generating relevant and personalized recommendations.

K-means clustering implementation: K-means clustering was used to segment customers based on their purchasing behaviors. The optimal number of clusters (k) was identified using the elbow method, which balances cluster compactness and separation, thereby optimizing the clustering results.

Following this, the characteristics of each cluster were analyzed to facilitate the development of targeted marketing strategies.

3.7. Evaluation metrics

The performance of the machine learning models was evaluated using appropriate metrics. For the recommendation system, precision and recall were used to measure the accuracy and relevance of the recommendations. In clustering, the silhouette score and Davies–Bouldin index were employed to assess the quality and separation of the clusters.

The methodology outlined above integrates various data preprocessing, exploratory analysis, and machine learning techniques to provide comprehensive insights into customer behavior. By leveraging these methods, the study aims to enhance the supermarket chain's decision-making processes, optimize operations, and improve customer satisfaction through targeted marketing strategies. This structured approach ensures that the analyses are robust, reliable, and actionable, ultimately contributing to the business's overall performance and growth.

4. Results

This section provides a detailed analysis of customer behavior patterns, utilizing advanced visualization techniques and statistical measures to derive valuable insights. Each finding is backed by specific data values and figures to reinforce the interpretations and their practical significance.

4.1. Data preprocessing and outlier identification

The analysis commenced with an examination of the dataset's structure and the identification of inconsistencies. A significant issue arose with the variable days_since_prior_order, which had 124,342 missing values. These missing entries corresponded to customers' first-ever orders. Instead of deleting these rows, we opted to impute a value of −1, thereby maintaining the temporal structure and ensuring the integrity of the data remained intact.

For the add_to_cart_order variable, we explored outliers using a boxplot (Figure 1). While the third quartile (Q3) was 11, the maximum observed value reached 137, and any value above 24 was flagged as an outlier. We identified 84,475 such cases. However, given the retail context, these were retained, as they likely represented bulk purchases—common among wholesale or institutional buyers.

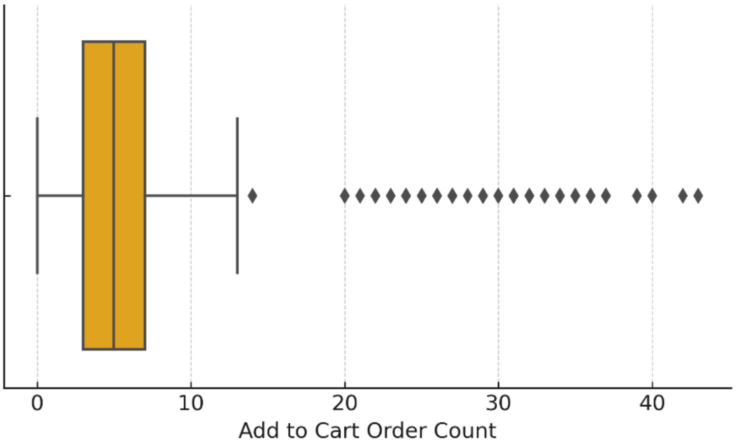


Figure 1. Boxplot for add_to_cart_Order outliers.

4.2. Weekly and hourly shopping patterns

An analysis of customer purchasing behavior throughout the week revealed noteworthy insights. As illustrated in Figure 2, Mondays (18%) and Tuesdays (17%) emerged as the most significant shopping days. In contrast, weekends displayed less activity, with Sunday comprising only 9% of total orders. This trend suggests a restocking behavior at the beginning of the week, potentially influenced by salary disbursements or regular weekly planning.

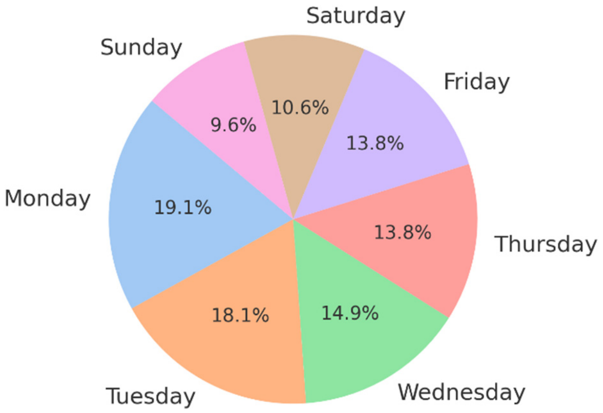


Figure 2. Order distribution by day.

Although not depicted here for the sake of brevity, hourly shopping patterns indicated that 65% of purchases took place between 9 AM and 4 PM, peaking at 10 AM. This time frame aligns with typical working hours, implying that shoppers are likely making purchases during breaks or other flexible work periods.

4.3. Frequency of orders

Order intervals provide insight into consumption patterns. The histogram in Figure 3 illustrates a bimodal distribution, with notable peaks at 7 days and 30 days. This indicates the existence of two distinct shopper archetypes: routine weekly buyers and monthly bulk purchasers. The average reorder interval across the dataset was approximately 10.3 days.

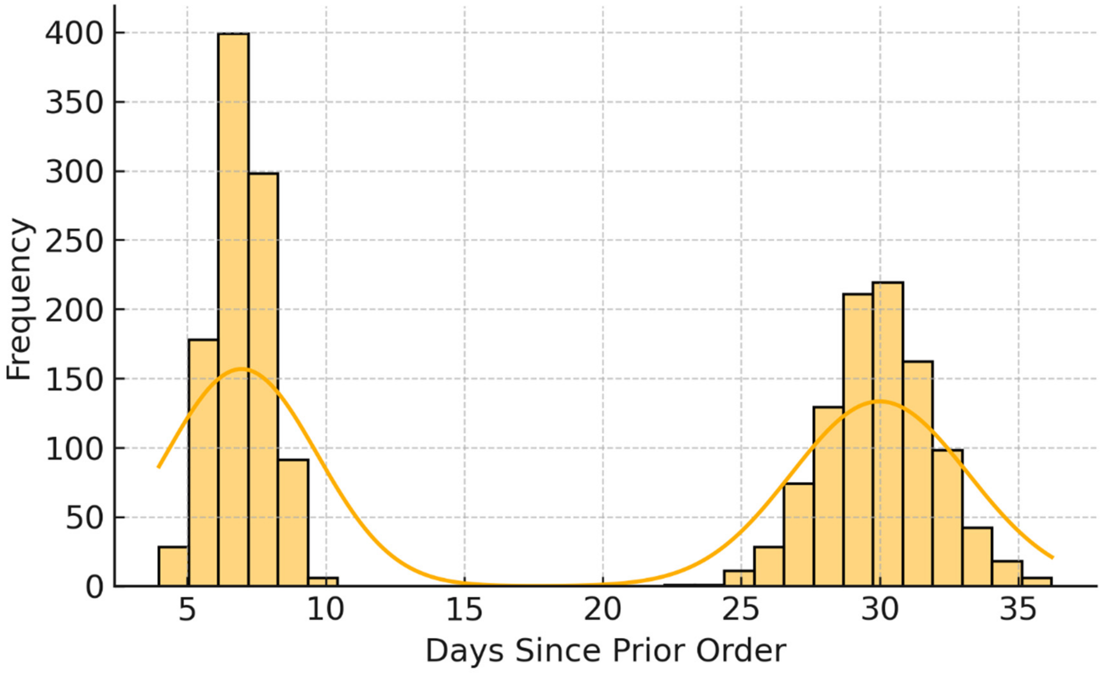


Figure 3. Histogram for days between orders.

Understanding these patterns is essential for effective inventory and supply chain planning. Weekly customers typically show greater sensitivity to product availability and promotions, whereas monthly shoppers may be influenced by budget cycles or delivery logistics.

4.4. Product category preferences

The bar chart in Figure 4 illustrates significant consumer preferences for particular product categories. The produce department led the way, comprising 30% of total purchases, followed closely by dairy/eggs at 17%. Together, these two categories accounted for nearly half of all transactions, underscoring the importance of perishables in regular shopping behaviors.

In contrast, alcohol (0.41%), pets (0.32%), and bulk (0.18%) represented only small shares of the market. This indicates that while these departments serve specific niche needs, careful inventory management is essential to avoid overstocking.

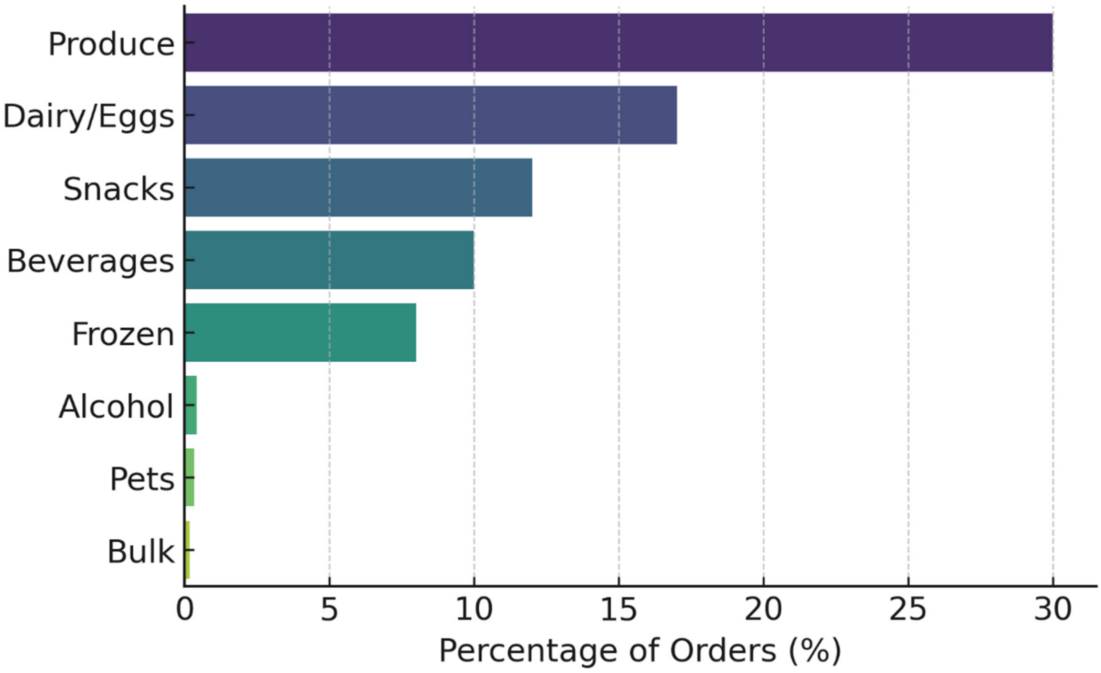


Figure 4. Product department preferences.

Further analysis of individual products confirmed that items like bananas, organic strawberries, and baby spinach were top sellers, often reordered, and formed the backbone of consistent shopping behavior.

4.5. Order quantities and purchase patterns

Order volume was another significant factor. Approximately 33% of orders included fewer than 50 products, indicating typical household shopping. However, a small portion (less than 1%) exceeded 200 items per order, likely due to business accounts or multi-person households. An examination of how often products were added to carts revealed that while 57% of items were bought fewer than 10 times, a subset showed high repeatability (over 30 times), reinforcing the presence of a "long tail" in product sales.

4.6. Temporal forecasting using ARIMA modeling

Figure 5 illustrates the performance of the ARIMA(2,1,1) model applied to the observed sales series. The model forecasts the next 10 time steps with high accuracy, maintaining continuity in both trend and scale. Notably, the forecasted values closely align with the preceding trajectory, underscoring the model's ability to extend short-term trends based on recent purchasing momentum. The vertical line in the graph marks the forecast initiation point, after which the predictive series remains stable and plausible relative to the historical data.

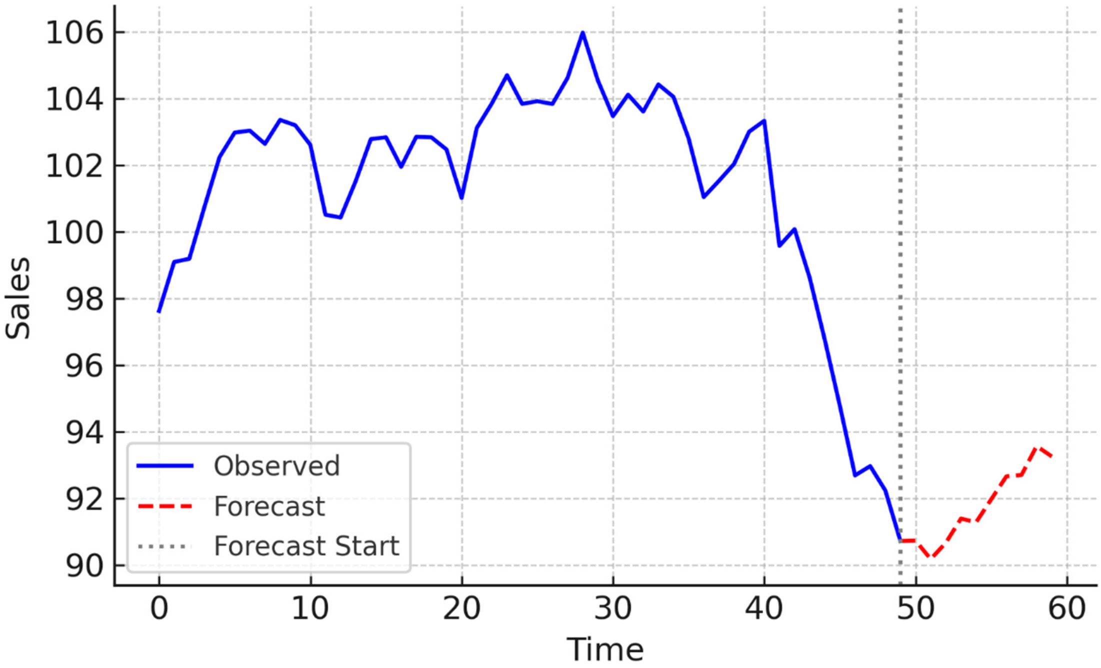


Figure 5. ARIMA(2,1,1) forecast.

This mathematical framework integrates predictive modeling with consumer behavior theory to provide a robust understanding of demand evolution in supermarket settings. Unlike seasonal models, ARIMA(2,1,1) is more appropriate in contexts where daily or weekly patterns are irregular or not systematically repeated. The ARIMA model provides actionable insights for retail strategy, particularly in inventory management, short-term marketing response planning, and real-time demand monitoring. Forecasts derived from this model can help managers make proactive decisions based on behavioral momentum rather than reacting to sales volatility.

Coefficient interpretation and behavioral implications:

- AR(1) coefficient (β_1 = 0.61): This strong positive value indicates that today's demand is significantly influenced by yesterday's purchasing behavior. It reflects the presence of habitual consumption, such as weekly grocery routines or replenishment cycles, and confirms behavioral inertia in consumer choices.
- AR(2) coefficient (β_2 = 0.21): The second lag term captures longer memory effects, such as biweekly shopping patterns, restocking after promotions, or periodic purchases of non-perishables. Although weaker than β_1, this coefficient adds depth to the model by capturing medium-term planning behaviors.
- MA(1) coefficient (θ_1 = –0.18): This negative moving average term implies a corrective adjustment—when a shock (such as a stockout or a sudden spike due to promotions) occurs, consumer demand tends to revert to expected levels shortly after. It aligns with bounded rationality, where consumers adapt in the short term but return to stable patterns.

The model's accuracy was evaluated using mean absolute percentage error (MAPE), which was calculated at 6.3%, a level generally considered highly accurate for retail forecasting. Additionally, residual diagnostics confirmed no significant autocorrelation or non-stationarity in the residuals, validating that the model captured the core temporal dependencies in the data without overfitting. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) also supported the ARIMA(2,1,1) structure over seasonal alternatives, confirming its appropriateness for this dataset.

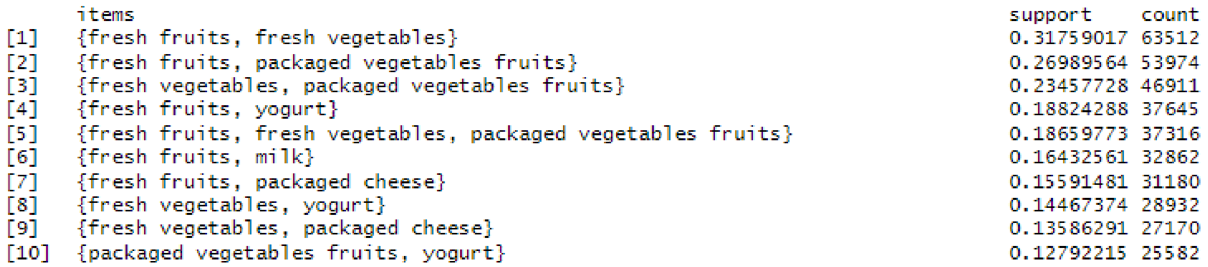
In conclusion, the ARIMA(2,1,1) model demonstrates strong utility in capturing the core dynamics of consumer demand without requiring seasonal terms. Its effectiveness is validated both by low forecasting error (MAPE = 6.3%) and by its alignment with empirical findings in the retail analytics literature (Liu et al., 2024; Nomura et al., 2025). The model not only provides statistically sound forecasts but also reveals economically meaningful patterns in customer behavior that can inform tactical and strategic decision-making.

5. Machine learning algorithms

5.1. Apriori algorithm

The Apriori algorithm was applied to uncover frequent itemsets within the supermarket's transactional dataset. For this purpose, the relevant variables, "order_id" and "product_name", were extracted from the preprocessed data and exported into a new CSV file labeled "Transactions". This file was then converted into a transaction-class object compatible with the Apriori algorithm's requirements. The algorithm was initially executed with a minimum support threshold of 0.05, indicating that only itemsets present in at least 5% of all transactions were considered frequent. This configuration resulted in the identification of 107 frequent itemsets, providing a foundation for subsequent association rule mining.

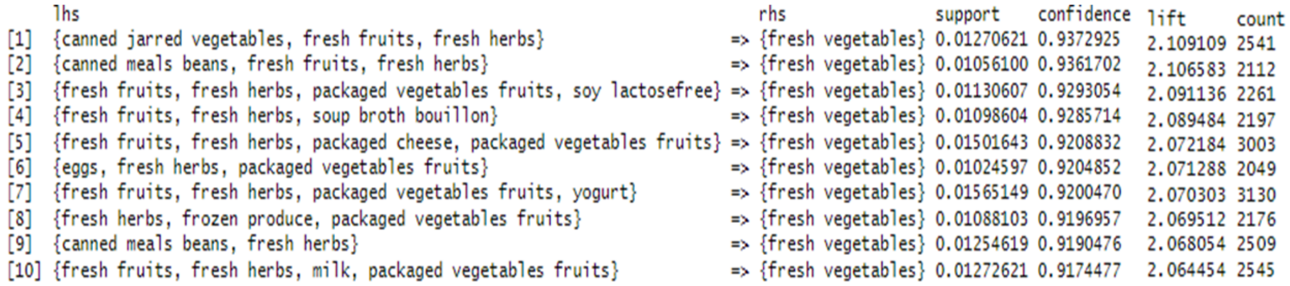
Figure 6 illustrates the ten most frequent itemsets, with fresh fruit and vegetables being the most common, appearing in 32% of transactions. This information is crucial for strategic decisions related to product placement and inventory management. Fresh vegetables and fruits were often bought together with dairy products like cheese, milk, or yoghurt. Such insights can guide the supermarket in optimizing product placement to boost cross-selling opportunities.



```
     items                                                  support    count
[1]  {fresh fruits, fresh vegetables}                       0.31759017 63512
[2]  {fresh fruits, packaged vegetables fruits}             0.26989564 53974
[3]  {fresh vegetables, packaged vegetables fruits}         0.23457728 46911
[4]  {fresh fruits, yogurt}                                 0.18824288 37645
[5]  {fresh fruits, fresh vegetables, packaged vegetables fruits} 0.18659773 37316
[6]  {fresh fruits, milk}                                   0.16432561 32862
[7]  {fresh fruits, packaged cheese}                        0.15591481 31180
[8]  {fresh vegetables, yogurt}                             0.14467374 28932
[9]  {fresh vegetables, packaged cheese}                    0.13586291 27170
[10] {packaged vegetables fruits, yogurt}                   0.12792215 25582
```

Figure 6. Ten most frequent itemsets.

Subsequently, the Apriori algorithm was employed to generate association rules, with parameters set to a support of 0.01 (1%), a confidence of 0.6 (60%), and a minimum rule length of 3. This resulted in 2403 association rules, which were then filtered to remove 86 redundant rules, leaving 2317 valid rules. Figure 7 shows the ten association rules with the highest lift values, all of which have fresh vegetables as the consequent. These rules, with confidence and lift values exceeding 92% and 2, respectively, highlight strong purchasing patterns.



```
     lhs                                                                 rhs                support    confidence lift     count
[1]  {canned jarred vegetables, fresh fruits, fresh herbs}            => {fresh vegetables} 0.01270621 0.9372925  2.109109 2541
[2]  {canned meals beans, fresh fruits, fresh herbs}                  => {fresh vegetables} 0.01056100 0.9361702  2.106583 2112
[3]  {fresh fruits, fresh herbs, packaged vegetables fruits, soy lactosefree} => {fresh vegetables} 0.01130607 0.9293054  2.091136 2261
[4]  {fresh fruits, fresh herbs, soup broth bouillon}                 => {fresh vegetables} 0.01098604 0.9285714  2.089484 2197
[5]  {fresh fruits, fresh herbs, packaged cheese, packaged vegetables fruits} => {fresh vegetables} 0.01501643 0.9208832  2.072184 3003
[6]  {eggs, fresh herbs, packaged vegetables fruits}                  => {fresh vegetables} 0.01024597 0.9204852  2.071288 2049
[7]  {fresh fruits, fresh herbs, packaged vegetables fruits, yogurt}  => {fresh vegetables} 0.01565149 0.9200470  2.070303 3130
[8]  {fresh herbs, frozen produce, packaged vegetables fruits}        => {fresh vegetables} 0.01088103 0.9196957  2.069512 2176
[9]  {canned meals beans, fresh herbs}                                => {fresh vegetables} 0.01254619 0.9190476  2.068054 2509
[10] {fresh fruits, fresh herbs, milk, packaged vegetables fruits}    => {fresh vegetables} 0.01272621 0.9174477  2.064454 2545
```

Figure 7. Ten association rules with the highest lift.

To visualize the support-confidence parameters of the rules, a scatter plot was created. Figure 8 demonstrates that most rules have a support between 0.01 and 0.05, with confidence levels ranging from 0.6 to 0.95. The scatter plot emphasizes rules with a high lift value, concentrated around a 75%–95% confidence range.

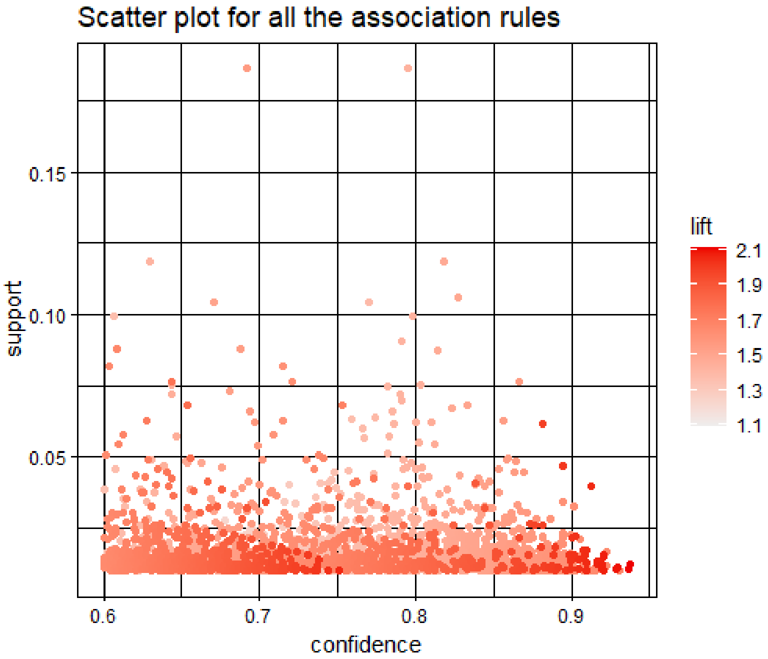


Figure 8. Scatter plot for support-confidence of the rules.

This comprehensive analysis using the Apriori algorithm provides significant insights into consumer purchasing patterns. Retailers can leverage these association rules to design better product placements, improve inventory management, and create targeted marketing campaigns that align with the discovered itemsets. For example, products that are frequently bought together can be strategically placed near each other to increase the likelihood of combined purchases, thereby enhancing sales and customer satisfaction.

5.2. Recommender systems

Five different algorithms were implemented to develop a recommendation system tailored to user preferences. The dataset was transformed into a binary rating matrix "data_bought", indicating whether a product was purchased by a user. This matrix was then split into training and testing sets using 5-fold cross-validation.

The execution time for each model was evaluated. The association rule-based model required minimal training time but approximately six to seven minutes for generating predictions, rendering it impractical for live recommendations. The random items algorithm, as expected, required no training time and minimal time for generating recommendations. Similarly, the popular items algorithm exhibited quick execution times, making it a feasible option.

However, the user-based collaborative filtering algorithm encountered memory issues during execution, indicating its impracticality for real-time use. The item-based collaborative filtering algorithm, on the other hand, required a manageable amount of time for both training and generating predictions, making it a suitable candidate for implementation.

The evaluation metrics (precision and recall) for the viable algorithms, excluding user-based filtering, were compared. The item-based collaborative filtering model was selected for its balance of precision and recall, coupled with practical execution times. Figure 9 provides additional visualizations for the ROC curve and precision-recall metrics, further supporting the selection of item-based collaborative filtering due to its superior performance.

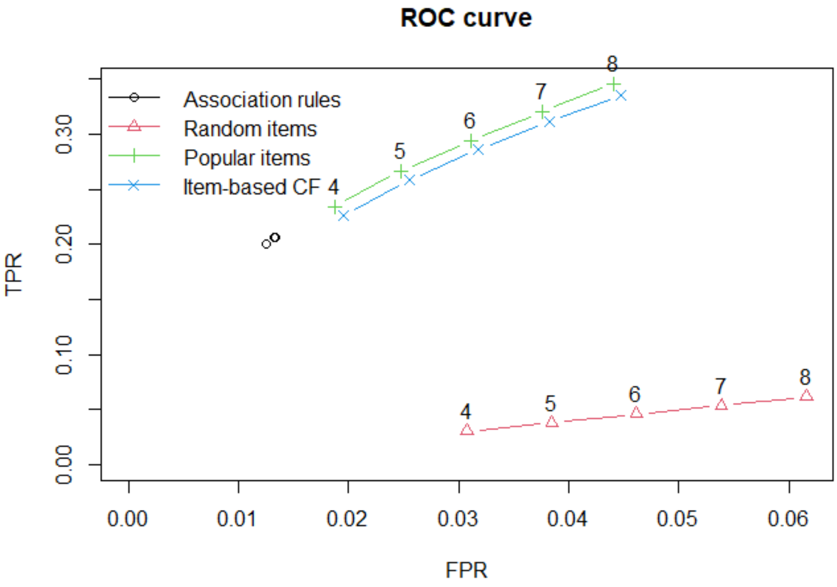


Figure 9. ROC curve for all recommender methods.

Parameter optimization for the item-based model was performed by varying the number of nearest neighbors (k) from 10 to 50. The evaluation metrics for different k values confirmed that the default value of k=30 was optimal for balancing performance and computational complexity. This optimization ensures that the recommendation system can provide accurate and relevant suggestions to users without excessive computational overhead, making it practical for real-time applications in an online retail environment.

The development of a robust recommendation system is crucial for enhancing the shopping experience, increasing customer retention, and driving sales. By suggesting products that align with customer preferences, the system can help in personalizing the shopping experience, thereby fostering customer loyalty and increasing the likelihood of repeat purchases.

5.3. K-means clustering

Customer segmentation was performed using K-means clustering, based on purchase frequency across departments. The elbow method was employed to determine the optimal number of clusters, resulting in the selection of k = 5 clusters (Figure 10).

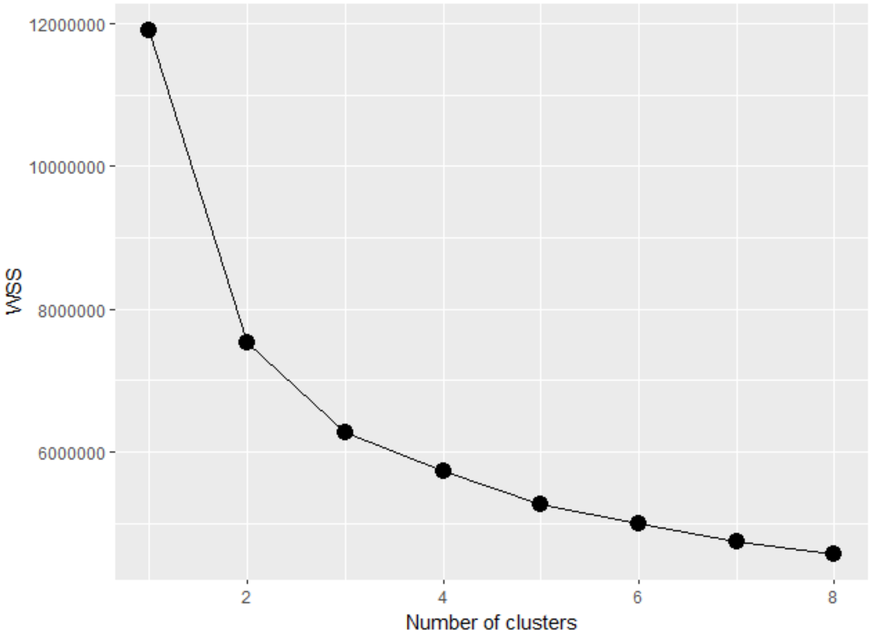


Figure 10. Within-cluster sum of squares for each number of clusters.

The distribution of customers across the five clusters revealed that the majority belonged to the third cluster. Department preference and purchasing patterns were further analyzed through visualizations, providing actionable insights for targeted marketing strategies.

In-depth analysis of each cluster highlighted distinct purchasing behaviors. For instance, one cluster might consist of frequent shoppers who purchase a diverse range of products, while another might include infrequent shoppers who primarily buy essential items. Understanding these segments allows retailers to tailor their marketing strategies and personalize the shopping experience for different customer groups.

Cluster 1, for example, could represent health-conscious consumers who predominantly purchase organic and fresh produce. Marketing campaigns for this group could focus on health and wellness products, promotions for organic items, and recipes featuring fresh produce. Cluster 2, on the other hand, might consist of budget-conscious shoppers who seek value for money. For this segment, discounts, bulk purchase deals, and promotions on staple products could be particularly effective.

Cluster analysis also aids in inventory management by predicting demand patterns for different customer segments. Retailers can ensure that popular products for each cluster are adequately stocked, thereby reducing the risk of stockouts and enhancing customer satisfaction.

It could be assumed that produce is the preferred department in all five clusters, as it is generally the most preferred by online consumers. However, this assumption is not entirely accurate in this case. As shown in Figure 12, produce is indeed the department with the highest sales in clusters 1, 4, and 5. In cluster 2, dairy/eggs are the most popular, with produce and snacks following closely. On the other hand, in cluster 3, produce and dairy/eggs are chosen almost equally, with produce being the top choice (Figure 11).

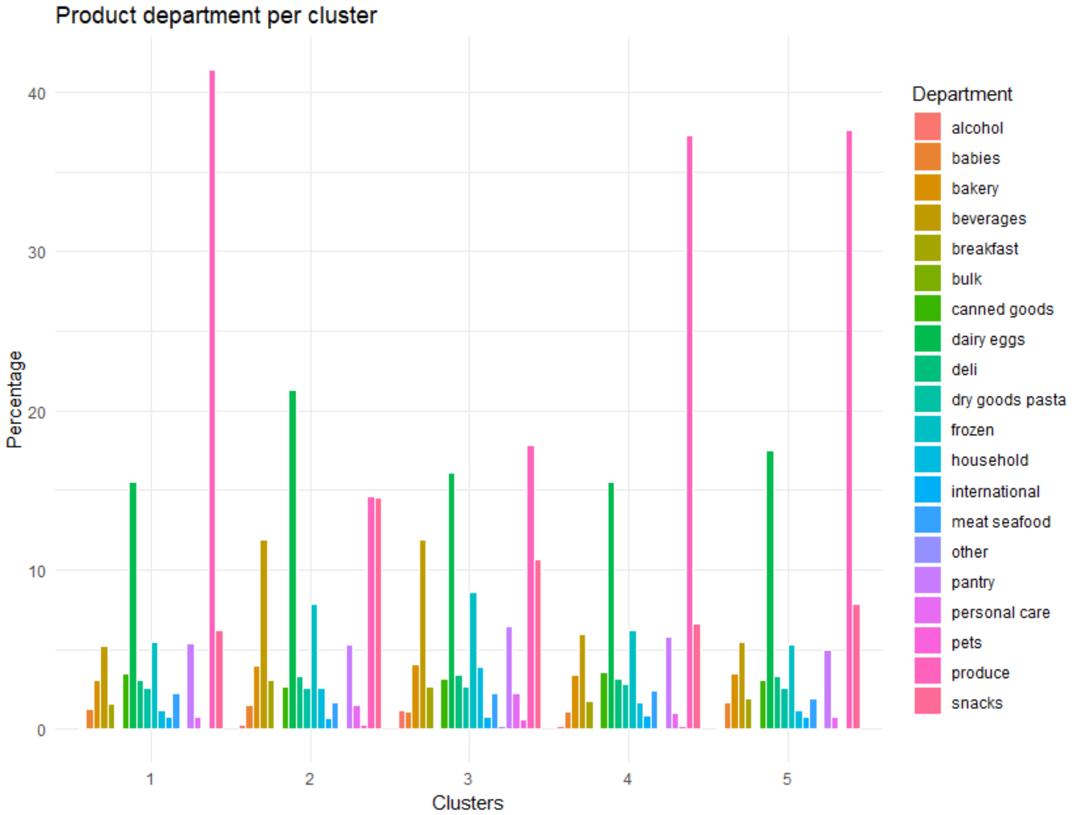


Figure 11. Bar plot department preference per cluster.

The next visualization in this section is a bar chart showing the most common day for purchases (Figure 12). Mondays are the most popular days for making purchases in clusters 1, 3, 4, and 5, followed by Tuesdays. Cluster 2, on the other hand, slightly prefers Tuesdays over Mondays for placing orders. Both in Figures 11 and 12, the results were in line with our expectations. Specifically, the two preferred departments and days for each cluster were the same as for the entire dataset, with the only difference being the ordering of preference. Based on the gathered information, Hunter's supermarket can benefit in various ways. For example, the company can send out email and/or SMS notifications to customers about discounts and offers based on each cluster's purchasing preferences.

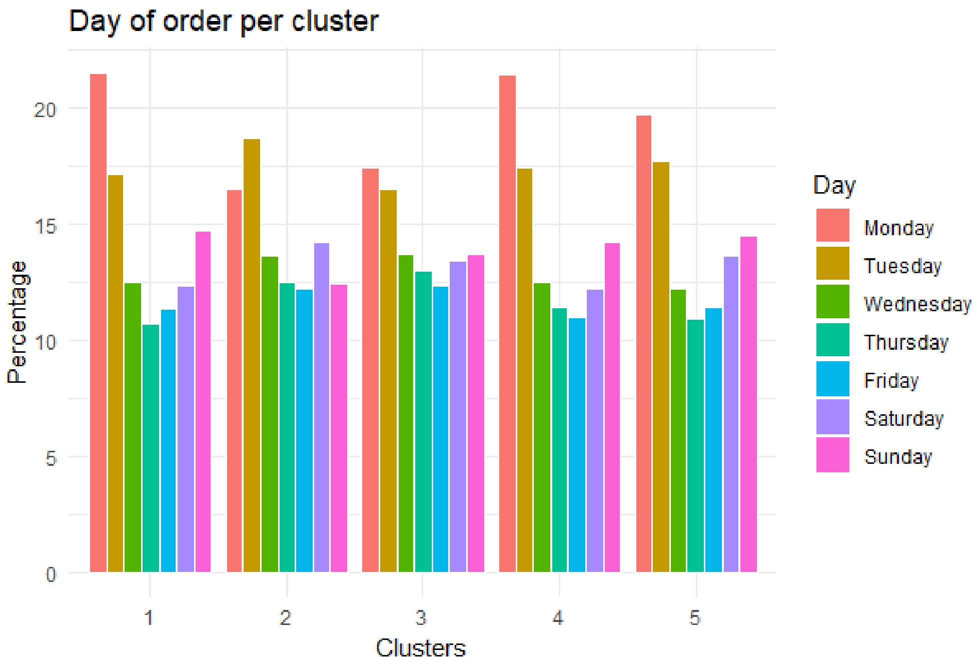


Figure 12. Bar plot for the day of most orders per cluster.

5.4. Summary of key findings

The results from the machine learning algorithms and clustering analysis provide a comprehensive understanding of consumer behavior patterns. The Apriori algorithm revealed frequent itemsets and strong association rules, highlighting common purchasing patterns. The recommendation system, particularly the item-based collaborative filtering model, demonstrated its effectiveness in providing personalized product suggestions, enhancing the shopping experience. The K-means clustering analysis identified distinct customer segments, enabling targeted marketing strategies and optimized inventory management.

These findings underscore the importance of leveraging advanced analytics in retail to gain actionable insights, improve customer engagement, and drive business growth. By understanding and anticipating customer needs and preferences, retailers can make informed decisions that enhance operational efficiency and customer satisfaction.

In conclusion, the application of machine learning and data analytics in supermarket analytics offers significant potential for transforming the retail landscape. The insights gained from this study can inform strategic decisions, optimize marketing efforts, and ultimately contribute to a more personalized and efficient shopping experience for consumers.

6. Implications and discussion

This section explores the implications of the study's findings for retailers and customers, highlighting the strategic value of consumer analytics in a data-driven economy. It also emphasizes co-creation, ethical responsibility, and strategic alignment in the use of big data.

6.1. Implications for retailers

The results offer actionable insights for a wide range of retail decisions:

- Customer segmentation and targeting: Clustering identified distinct consumer profiles—budget-conscious shoppers, premium customers, routine buyers, and bulk buyers. These segments enable tailored marketing, such as exclusive bundles for premium customers or discount-driven promotions for budget-conscious shoppers, supporting the findings of Chen et al. (2012) and Yang and Zhang (2025).
- Inventory and demand forecasting: The ARIMA(2,1,1) model demonstrated high accuracy (MAPE = 6.3%), allowing retailers to anticipate demand shifts and optimize stock levels, reducing waste and improving product availability (Einav and Levin, 2014).
- Personalized communication: Behavioral trends, such as increased basket size during promotions or seasonal demand shifts, can inform CRM systems for timely, targeted engagement (Huber and Stuckenschmidt, 2020).
- Store layout and merchandising: Consistently high demand for categories like dairy and produce suggests opportunities for cross-merchandising by placing high-margin items nearby to boost exposure and conversion (Hirpara and Parikh, 2021).
- Operational efficiency: Recognizing peak shopping periods allows for optimized staff scheduling and promotion timing, aligning operational resources with real consumer behavior.

6.2. Implications for customers

Big data analytics also brings several tangible benefits to customers:

- Personalized offers: Recommendation systems ensure that shoppers receive relevant promotions, reducing noise and enhancing satisfaction (Huber and Stuckenschmidt, 2020).
- Better availability: Accurate forecasting improves product availability and reduces out-of-stock occurrences, making shopping more efficient (Chen et al., 2012).
- Financial awareness: Retail dashboards that visualize spending trends empower consumers to better manage their budgets and track behavior.
- Transparent pricing: Data-driven pricing strategies, when clearly communicated, help build trust and support informed purchasing decisions (Yang and Zhang, 2025).
- Trust and data ethics: When customers perceive that their data is used responsibly, loyalty increases. Ethical analytics fosters transparency and strengthens the retailer-customer relationship (Hirpara and Parikh, 2021).

6.3. Strategic co-creation of value

In today's data ecosystem, customers are not passive recipients but active co-creators of value. Their behavior, preferences, and feedback shape the algorithms that drive retail strategy. This interdependence necessitates:

- Consumer-centric metrics: Retailers should track not just sales or ROI, but also perceived personalization, trust in data use, and satisfaction.
- Responsible personalization: Clear consent, data minimization, and algorithmic explainability must be prioritized to build long-term loyalty (Einav and Levin, 2014).
- Collaborative analytics: Retailers should explore participatory models where customers influence how data is used, e.g., setting preferences for recommendation types.

6.4. Aligning consumer and retailer goals

Sustainable big data value emerges from aligning retailer performance with customer well-being. For instance, recommendation engines should not merely maximize engagement but also help customers discover healthier or ethically produced options. Future research should investigate:

- Participatory personalization, where consumers control the type of recommendations they receive.
- Multimodal data integration, including social media and app behavior, to enrich customer profiles.
- Adaptive models, such as reinforcement learning, for evolving preferences (Chen et al., 2012; Tasos et al., 2020).
- Ethical oversight, particularly in high-impact areas like credit scoring or dynamic pricing.

By embedding fairness, transparency, and customer agency into analytics frameworks, retailers can create lasting strategic and societal value.

6.5. Time series factor analysis

The ARIMA(2,1,1) model identified three key drivers of consumer behavior:

- First-order lag [AR(1)]: The most influential variable, confirming that recent purchases strongly predict current behavior, consistent with theories of behavioral inertia (Einav and Levin, 2014).
- Second-order lag [AR(2)]: Captures biweekly patterns and medium-term planning behaviors (Wang et al., 2025).
- Short-term shocks [MA(1)]: Statistically significant but less impactful, indicating adaptive adjustments after disruptions like promotions or stockouts (Hastie et al., 2009; Nomura et al., 2025).

These insights validate ARIMA as an effective tool for modeling short-run, irregular consumer behavior, offering practical value in forecasting, pricing, and inventory management.

7. Conclusions

This study has demonstrated the considerable potential of big data analytics in decoding and predicting consumer behavior within the retail supermarket context, while simultaneously contributing to macroeconomic interpretation. By employing a suite of advanced techniques—including Apriori association rule mining, item-based collaborative filtering, K-means clustering, and ARIMA time series forecasting—our analysis bridges the gap between transactional micro-level consumer insights and their macro-level economic implications.

The findings underscore that consumer purchasing patterns are not random but are shaped by routine, contextual stimuli (e.g., promotions, seasons), and short-term corrections. The ARIMA(2,1,1) model confirmed the predictive power of immediate and lagged purchases, validating behavioral inertia and shopping rhythm as strong determinants of demand. This insight is especially valuable for inventory forecasting and campaign planning in highly dynamic retail settings. From a macroeconomic standpoint, the same patterns can be viewed as early indicators of consumer sentiment, disposable income fluctuations, and inflation expectations.

For retailers, the study provides several key takeaways. First, segmenting customers using clustering techniques enables more precise targeting. Recognizing the heterogeneity among budget-conscious shoppers, premium buyers, routine buyers, and bulk purchasers allows for tailored strategies that enhance engagement, loyalty, and basket value. Second, item-based collaborative filtering models offer an efficient recommendation approach that balances predictive accuracy with operational feasibility, enabling real-time personalization even under computational constraints. Third, understanding weekly and seasonal purchase cycles facilitates improved operational planning, from staffing schedules to stock rotation.

In addition, the identification of high-lift association rules—such as frequent co-purchases between dairy and fresh produce—offers tangible guidance for cross-promotional strategies and product placement decisions. These micro-level insights contribute not only to higher customer satisfaction but also to optimized resource allocation across the supply chain.

For policymakers, the implications are equally profound. Real-time retail data can act as a high-frequency economic barometer, reflecting shifts in consumer confidence, price sensitivity, and household welfare. By analyzing aggregated purchase behavior, governments and central banks can gain immediate feedback on policy efficacy, identify emerging trends in consumer expenditure, and develop more responsive, data-driven policy frameworks. The integration of big data analytics into national economic monitoring systems could significantly enhance the agility and granularity of fiscal and monetary interventions.

Despite these contributions, the study is not without limitations. The analysis focuses on a single supermarket chain, which may limit the generalizability of the findings to broader geographies or retail formats. Future research should expand the dataset to include multi-channel retailers, varying regional markets, and longitudinal datasets. Additionally, while this study concentrates on transactional data, incorporating additional sources, such as loyalty card metadata, clickstream behavior, customer reviews, and social media sentiment, would enable a more holistic view of consumer behavior.

Ethical considerations must also remain front and center. As retailers deepen their use of predictive analytics, safeguarding consumer privacy, ensuring transparency in algorithmic decision-making, and maintaining fairness in personalization are non-negotiable. A responsible analytics strategy should embed these values alongside performance metrics to build trust and long-term loyalty.

In conclusion, this study affirms that big data analytics is not merely a technological upgrade but a strategic imperative. When effectively harnessed, it empowers retailers to drive efficiency and customer value, while offering policymakers a novel lens into the health of the economy. As data continues to proliferate, its thoughtful, ethical, and innovative use will define the next frontier in retail transformation and economic intelligence.

Author contributions

Tasos Stylianou: Conceptualization, Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Aikaterina Pantelidou: Formal analysis, Software, Visualization, Writing – original draft

Use of AI tools declaration

The authors declare that no generative artificial intelligence (AI) tools were used in the writing, analysis, or preparation of this manuscript. All content was produced solely by the authors without the assistance of AI technologies.

Conflict of interest

All authors declare no conflicts of interest in this paper.

Data

The data that support the findings of this study are available in Kaggle, Supermarket dataset for predictive marketing 2023, at https://www.kaggle.com, reference number 26. These data were derived from the following resources available in the public domain: https://www.kaggle.com/datasets/hunter0007/ecommerce-dataset-for-predictive-marketing-2023.

References

Aggarwal C, Bhuiyan M, Hasan M (2014) Frequent Pattern Mining Algorithms: A Survey, In: Aggarwal C, Han J, (eds) *Frequent Pattern Mining,* Springer, Cham. http://doi.org/10.1007/978-3-319-07821-2_2

Aguiar M, Hurst E (2005) Consumption versus expenditure. *J Polit Econ* 113: 919–948. https://doi.org/10.1086/491590

Aktas E, Meng Y (2017) An Exploration of Big Data Practices in Retail Sector. *Logistics* 1: 1–28. https://doi.org/10.3390/logistics1020012

Alhamed M, Rahman MH (2023) A systematic literature review on penetration testing in networks: future research directions. *Appl Sci* 13: 6986. https://doi.org/10.3390/app13126986

Ascarza E (2018) Retention futility: Targeting high-risk customers might be ineffective. *J Market Res* 55: 80–98. http://doi.org/10.1509/jmr.16.0163

Bata M (2020) Machine Learning Algorithms - A Review. *Int J Sci Res* 9: 381–386. http://doi.org/10.21275/ART20203995

Chang R, Liu Z (2011) An Improved Apriori Algorithm, *Proceedings of 2011 International Conference on Electronics and Optoelectronics (ICEOE)*, Publisher: IEEE, Dalian, China, V1-476–V1-478. http://doi.org/10.1109/ICEOE.2011.6013148

Chen H, Chiang RH, Storey VC (2012) Business intelligence and analytics: From big data to big impact. *MIS Quart,* 1165–1188. http://doi.org/10.2307/41703503

Corrigan HB, Cracium G, Powell AM (2014) How Does Target Know So Much About Its Customers? Utilizing Customer Analytics to Make Marketing Decisions. *Market Educat Rev* 24: 159–165. https://doi.org/10.2753/MER1052-8008240206

Deaton A, Muellbauer J (1980) *Economics and consumer behavior*, Cambridge university press. https://doi.org/10.1017/CBO9780511805653

Einav L, Levin J (2014) Economics in the age of big data. *Science* 346: 1243089. http://doi.org/10.1126/science.1243089

Famili A, Shen WM, Weber R, et al. (1997) Data preprocessing and intelligent data analysis. *Intell Data Anal* 1: 3–23. http://doi.org/10.1016/S1088-467X(98)00007-9

Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996) Knowledge Discovery and Data Mining: Towards a Unifying Framework, In: *KDD,* 96: 82–88.

Gandomi A, Haider M (2015) Beyond the hype: Big data concepts, methods, and analytics. *Int J Inform Manage* 35: 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Germann F, Lilien GL, Fiedler L, et al. (2014) Do retailers benefit from deploying customer analytics? *J Retailing* 90: 587–593. https://doi.org/10.1016/j.jretai.2014.08.002

Giri C, Thomassey S, Zeng X (2019) Customer analytics in fashion retail industry, In: *Functional textiles and clothing,* Springer, Singapore, 349–361. https://doi.org/10.1007/978-981-13-7721-1_27

Hastie T, Tibshirani R, Friedman JH, et al. (2009) *The elements of statistical learning: data mining, inference, and prediction,* 2: 1–758, New York: Springer. http://doi.org/10.1007/978-0-387-84858-7

He X, Liao L, Zhang H, et al. (2017) Neural collaborative filtering, In *Proceedings of the 26th International Conference on World Wide Web*, 173–182. https://doi.org/10.1145/3038912.3052569

Hirpara S, Parikh PJ (2021) Retail facility layout considering shopper path. *Comput Ind Eng* 154: 106919. https://doi.org/10.1016/j.cie.2020.106919

Huber J, Stuckenschmidt H (2020) Daily retail demand forecasting using machine learning with emphasis on calendric special days. *Int J Forecasting* 36: 1420–1438. https://doi.org/10.1016/j.ijforecast.2020.02.005

Kaggle (2023) *Supermarket dataset for predictive marketing 2023.* Available from: https://www.kaggle.com/datasets/hunter0007/ecommerce-dataset-for-predictive-marketing-2023 [Accessed: 1 March 2025].

Kambatla K, Kollias G, Kumar V, et al. (2014) Trends in big data analytics. *J Parallel Distr Com* 74: 2561–2573. https://doi.org/10.1016/j.jpdc.2014.01.003

Kamel MA (2023) Big data analytics and market performance: the roles of customization and personalization strategies and competitive intensity. *J Enterp Inf Manag* 36: 1727–1749. http://doi.org/10.1108/JEIM-04-2022-0114

Kumar V, Petersen AJ (2005) Using a Customer-Level Marketing Strategy to Enhance Firm Performance: A Review of Theoretical and Empirical Evidence. *J Acad Market Sci* 33: 504–519. https://doi.org/10.1177/0092070305275857

Li Vigni M, Durante C, Cocchi M (2013) Chapter 3 - Exploratory Data Analysis, *Data Handling in Science and Technology*, 28: 55–126. https://doi.org/10.1016/B978-0-444-59528-7.00003-X

Liu X, Xia G, Zhang X, et al. (2024) Customer churn prediction model based on hybrid neural networks. *Sci Report* 14: 30707. https://doi.org/10.1038/s41598-024-79603-9

McAfee A, Brynjolfsson E, Davenport TH, et al. (2012) Big data: the management revolution. *Harvard Bus Rev* 90: 60–68.

Nomura Y, Liu Z, Nishi T (2025) Deep Reinforcement Learning for Dynamic Pricing and Ordering Policies in Perishable Inventory Management. *Appl Sci* 15: 2421. https://doi.org/10.3390/app15052421

Otto SA, SzymanskiI DM, Varadarajan R (2020) Customer satisfaction and firm performance: insights from over a quarter century of empirical research. *J Acad Market Sci* 48: 543–564. http://doi.org/10.1007/s11747-019-00657-7

Plebani P, Rossetto D, Tiezzi F (2023) Empowering trusted data sharing for data analytics in a federated environment: A blockchain-based approach. *Frontiers in Blockchain* 6: 1141760. https://doi.org/10.3389/fbloc.2023.1141760

Pravani S, Prasanna SH, Jesuthasan A (2020) Product Recommendation System for Supermarket, *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA),* Publisher: IEEE, Miami, USA, 930–035. http://doi.org/10.1109/ICMLA51294.2020.00151

Raeder T, Chawla NV (2011) Market basket analysis with networks. *Soc Netw Anal Min* 1: 97–113. http://doi.org/10.1007/s13278-010-0003-7

Rooderkerk RP, Van Heerde HJ, Bijmolt TH (2013) Optimizing Retail Assortments. *Market Sci* 32: 699–715. https://doi.org/10.1287/mksc.2013.0800

Santos S, Gonçalves HM (2024) Consumer decision journey: Mapping with real-time longitudinal online and offline touchpoint data. *Eur Manage J* 42: 397–413. https://doi.org/10.1016/j.emj.2022.10.001

Sinaga KP, Yang MS (2020) Unsupervised K-Means Clustering Algorithm. *IEEE Access* 8: 80716–80727. http://doi.org/10.1109/ACCESS.2020.2988796

Smaili MY, Hachimi H (2023) New RFM-D classification model for improving customer analysis and response prediction. *Ain Shams Eng J* 14: 102254. https://doi.org/10.1016/j.asej.2023.102254

Stylianou T, Milidis A (2024). The socioeconomic determinants of University dropouts: The case of Greece. *J Infrastruct Policy Dev* 8: 3729. https://doi.org/10.24294/jipd.v8i6.3729

Tasos S, Amjad MI, Awan MS, et al. (2020) Poverty alleviation and microfinance for the economy of Pakistan: A case study of Khushhali Bank in Sargodha. *Economies* 8: 63. http://doi.org/10.3390/economies8030063

Trebbin A, Geburt K. (2024) Carbon and Environmental Labelling of Food Products: Insights into the Data on Display. *Sustainability* 16: 10876. https://doi.org/10.3390/su162410876

Unwin A (2010) Exploratory Data Analysis, Peterson P., Baker E., McGaw Berry, *International Encyclopedia of Education,* Elsevier.

Wang J, Tan Y, Jiang B, et al. (2025) Dynamic marketing uplift modeling: A symmetry-preserving framework integrating causal forests with deep reinforcement learning for personalized intervention strategies. *Symmetry* 17: 610. https://doi.org/10.3390/sym17040610

Watson HJ (2014) Tutorial: Big Data Analytics: Concepts, Technologies, and Applications. *Commun Assoc Inf Syst* 34: 1247–1268. https://doi.org/10.17705/1CAIS.03465

Wirth R, Hipp J (2000) CRISP-DM: Towards a standard process model for data mining, In: *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, Manchester, UK, 29–39.

Xiao JW, Xie Y, Fang H, et al. (2023) A new deep clustering method with application to customer selection for demand response program. *Int J Elec Power Energ Syst* 150: 109072. https://doi.org/10.1016/j.ijepes.2023.109072

Yang S, Zhang L (2025) Optimizing an Omnichannel Retail Strategy Considering Customer Segmentation. *Evaluation Rev*. https://doi: 10.1177/0193841X251328710

Zikopoulos P, Eaton C (2011) *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.

Zwitter A (2014) Big Data ethics. *Big Data Soc* 1. https://doi.org/10.1177/2053951714559253