Decision Analytics Journal 16 (2025) 100600

Contents lists available at ScienceDirect

Decision Analytics Journal

journal homepage: www.elsevier.com/locate/dajour

A machine learning approach to consumer behavior in supermarket analytics

Tasos Stylianou [a,*], Aikaterina Pantelidou [b]

[a] Department of Economics, University of Macedonia, Dimitriou Poliorkitou 65, PC 54633 Thessaloniki, Greece
[b] School of Computing, Mediterranean College, Thessaloniki, Greece

ARTICLE INFO

Dataset link: https://www.kaggle.com, https://www.kaggle.com/datasets/hunter0007/ecommerce-dataset-for-predictive-marketing-2023

JEL classification:
C45
C81
D12
L81

Keywords:
Consumer analytics
Machine learning
Data mining
Market trends
Predictive analytics
Big data

ABSTRACT

The rapid advancement of Big Data technologies has significantly influenced multiple sectors, with the retail industry being a key impact area. This study explores the relationship between Big Data analytics and consumer behavior, focusing on supermarket transaction data to extract macroeconomic insights. Using a dataset of over two million records from a multinational supermarket chain, the research employs a suite of advanced analytical techniques, including the Apriori algorithm for association rule mining, K-Means clustering for customer segmentation, collaborative filtering for recommendation systems, and AutoRegressive Integrated Moving Average (ARIMA) for time series forecasting. The study identifies purchasing patterns, segment-specific preferences, and temporal shopping behaviors. The results reveal distinct associations among frequently co-purchased products, clear segmentations based on shopping habits, and predictive trends in consumer demand, offering valuable input for marketing, inventory management, and policy-making. Beyond the operational insights, this study highlights the potential of transactional data to reflect broader economic shifts, such as changes in consumption patterns during periods of economic uncertainty, thus linking consumer micro-behaviors to macroeconomic indicators. The novelty of this work lies in its integration of multiple machine learning techniques within a unified framework that connects retail analytics to economic policymaking, thereby extending the application of Big Data from commercial strategy to public economics.

1. Introduction

In the digital era, the unprecedented generation and accumulation of data have propelled Big Data to the forefront of industrial innovation. In retail, extensive datasets from transactions, loyalty programs, and online platforms provide vital information on consumer behavior. This research explores Big Data and consumer behavior via supermarket analytics, aiming to uncover significant patterns with macroeconomic relevance [1].

Big Data, characterized by volume, velocity, variety, veracity, and value, surpasses traditional data processing capacities. Supermarket data, derived from diverse sources such as point-of-sale systems and loyalty programs, enables analysis of consumer behavior, informing business strategies and economic policies.

Understanding consumer behavior involves examining purchasing decisions and responses to market stimuli. Insights gained optimize product offerings, refine marketing strategies, and inform broader economic policymaking. Supermarket analytics employs advanced techniques such as machine learning and data mining to predict trends and enhance operational efficiency.

Studying consumer behavior in the retail industry entails delving into the intricacies of how individuals make purchasing decisions and the various factors that shape these decisions. This involves understanding consumer preferences for specific products, analyzing their responses to marketing strategies, and evaluating the influence of economic conditions on their purchasing power [1,2]. Gaining insight into consumer behavior is crucial for retailers as it allows them to optimize their product offerings, tailor their marketing initiatives to individual preferences, and enhance overall customer satisfaction. Furthermore, these insights have broader implications as they can be used to inform economic policies and strategies to foster economic growth and stability.

Supermarket analytics involves the comprehensive utilization of advanced data analytics techniques to meticulously process a wide range of supermarket data. This encompasses an in-depth analysis of various aspects such as sales data, inventory levels, customer demographics, and purchasing patterns to derive valuable and actionable insights. The overarching objectives of supermarket analytics include the optimization of operational efficiency, the augmentation of customer experiences, and the facilitation of sales growth through strategic decision-making informed by data-driven insights. The implementation of sophisticated techniques such as machine learning, data mining, and rigorous statistical analysis plays a pivotal role in conducting

* Corresponding author.
E-mail addresses: tasosstylianou@gmail.com (T. Stylianou), a.c.pantelidou@mc-class.gr (A. Pantelidou).

https://doi.org/10.1016/j.dajour.2025.100600
Received 26 January 2025; Received in revised form 25 May 2025; Accepted 25 June 2025
Available online 2 July 2025

thorough processing and analysis of supermarket data. The gathered insights serve as a foundation for predicting future sales trends, fine-tuning inventory management processes, and crafting tailored marketing strategies that effectively resonate with consumer preferences and behavior.

This study addresses supermarkets' challenges in understanding customer behavior using Big Data analytics, examining transaction data to provide insights into economic influences on consumer decisions. Through advanced techniques like machine learning, time series analysis, clustering, and association rule mining using the Apriori algorithm, this research aims to uncover intricate patterns and trends in consumer purchasing behavior.

Although this study leverages association rule mining to extract frequent itemsets and purchasing patterns, it complements this with clustering and predictive analytics to provide a more holistic and dynamic understanding of customer behavior. This multifaceted approach enables not only the discovery of static associations between products but also the segmentation of consumers and the prediction of future trends.

1.1. Big Data and macroeconomic perspectives

From a macroeconomic standpoint, the examination of Big Data derived from supermarket transactions can yield valuable insights into overarching economic patterns. Specifically, shifts in consumer expenditure patterns can function as barometers of economic vitality. During periods of economic recession, consumers may modify their purchase behaviors from non-essential items to essential commodities, thereby mirroring wider economic circumstances [3,4]. Conversely, heightened expenditure on non-essential items may be indicative of economic recuperation and burgeoning consumer confidence. Through the comprehensive analysis of supermarket data on a substantial scale, economists and policymakers can promptly ascertain the state of the economy, thereby facilitating the implementation of more adaptive and efficacious economic strategies.

1.2. Objectives of the study

This study seeks to investigate the macroeconomic implications of Big Data analytics in comprehending consumer behavior using supermarket data. The specific objectives include analyzing consumer purchasing patterns and identifying trends over time, examining the impact of economic conditions on consumer behavior in the supermarket sector, evaluating the effectiveness of various Big Data analytics techniques in predicting consumer behavior, and providing recommendations for retailers and policymakers based on the findings.

The research makes use of a thorough dataset collected from a supermarket chain, consisting of transaction records, customer demographics, and product details. Advanced analytics methods such as machine learning algorithms, time series analysis, and clustering are utilized to process and scrutinize the data. The examination concentrates on pinpointing important trends in consumer behavior, comprehending the driving forces behind these trends, and evaluating their broader economic implications.

Overall, the Objectives of the Study are:

- Analyze consumer purchasing patterns and identify trends.
- Examine economic conditions on supermarket consumer behavior.
- Evaluate Big Data analytics techniques' effectiveness in predicting behavior.
- Offer recommendations for retailers and policymakers.

1.3. Significance of the study

For retailers, Big Data analytics enhances strategy formulation, customer satisfaction, and competitiveness. For policymakers, insights into consumer behavior support economic policy development. This study contributes empirical evidence to the growing literature on Big Data's macroeconomic relevance.

1.3.1. Novelty and contribution

While previous studies have primarily focused on leveraging Big Data to optimize marketing and operational strategies within the retail sector, this study offers a novel contribution by extending supermarket analytics to a macroeconomic perspective. By systematically integrating association rule mining, clustering algorithms, and recommendation systems, the research not only enhances business decision-making but also provides a predictive framework for monitoring economic trends through shifts in consumer behavior. This multidimensional approach bridges the gap between micro-level consumer insights and macro-level economic analysis.

This study addresses the critical issues that supermarkets face in understanding customer behavior by leveraging Big Data analytics. Through advanced techniques like machine learning and time series analysis, this research aims to uncover intricate patterns and trends in consumer purchasing behavior. By examining transaction data from a supermarket chain, the study provides insights into how economic factors influence consumer decisions, how purchasing patterns evolve, and how these insights can be used to enhance business strategies and inform economic policies. The behavior of customers is identified through a combination of clustering, predictive analytics, and association rule mining, providing a comprehensive understanding of consumer preferences and trends.

1.4. Methodological framework

This study applies a combination of advanced data analytics techniques to uncover consumer behavior patterns within supermarket transactional data. Machine learning models were employed, including the Apriori algorithm for frequent itemset mining and association rule mining, collaborative filtering for the development of a recommendation system, and K-Means clustering for customer segmentation. Time series analysis using the ARIMA model was applied to predict future sales trends based on historical purchasing data. Data preprocessing included handling missing values, treating outliers, and standardizing formats to ensure data quality and consistency. These methods were selected for their proven effectiveness in uncovering hidden patterns, enabling predictive modeling, and providing actionable insights in retail analytics. We did not only use association rules because behavior is dynamic, and association rules alone capture static relations, not evolving preferences.

This study adopts a structured framework encompassing:

- Data Collection
- Data Preprocessing
- Exploratory Data Analysis
- Machine Learning Model Development (Apriori, Recommender Systems, K-Means Clustering)
- Model Evaluation
- Result Interpretation

This workflow bridges raw transactional data with actionable consumer insights (see Fig. A).

In the following sections, this paper will examine the pertinent literature on Big Data and consumer behavior, outline the methodology employed for data analysis, present the results, and delve into their implications. The aim is to offer a comprehensive grasp of how Big Data analytics can be utilized to gain a macroeconomic perspective on consumer behavior, ultimately benefiting both the retail industry and the broader economy.

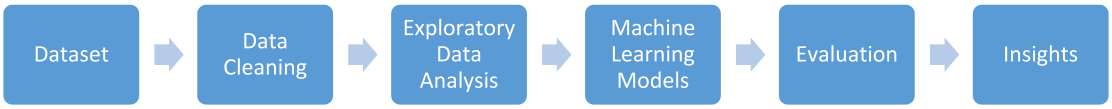

Fig. A. Research workflow for Big Data supermarket analytics.

Table 1
Summary of major studies on big data analytics and consumer behavior.

Authors	Methodology	Main findings	Limitations
Gandomi & Haider (2015)	Literature Review	Defined the 5Vs of Big Data and discussed analytics techniques	Lacked empirical focus
Manrik (2020)	Case Studies	Highlighted importance of predictive analytics in retail	Focused on developed countries
Germann et al. (2014)	Empirical Analysis	Found that customer analytics improve firm performance	Limited sectoral analysis
Singh & Thakur (2022)	Survey Research	Identified key challenges and opportunities in Big Data retail applications	Regional bias in sample
Zhao et al. (2021)	Real-time ML Models	Developed predictive models for consumer behavior	Need for deeper model validation

2. Literature review

2.1. Big Data and analytics

When considering Big Data (BD), the initial thought often revolves around the vast volume of data. However, BD encompasses more than just volume; it is characterized by high volume, velocity, variety, veracity, and value. Handling these elements requires sophisticated technologies and techniques to efficiently capture, store, and analyze the data, improving decision-making and identifying new patterns [5–7]. Recent studies emphasize how BD impacts industries such as healthcare, finance, and particularly retail [8,9].

Recent literature underscores the increasing sophistication of Big Data analytics in retail. For instance, **Bakare et al. (2024)** [10] examined data governance challenges in GDPR-compliant consumer analytics, emphasizing ethical use of purchasing data. **Singh and Thakur (2022)** [8] provided a comprehensive review of Big Data in retail, noting the practical applications of **Apriori and FP-Growth** for consumer profiling. Similarly, **Zhao et al. (2021)** [9] demonstrated the use of real-time machine learning models – integrating Apriori rules – for predicting in-store purchase behavior. **Tran and Hoang (2023)** [11] extended customer segmentation by integrating K-Means clustering and Apriori for e-commerce personalization, while **Chen et al. (2022)** [12] proposed a hybrid model combining collaborative filtering and Apriori for product recommendations. Additionally, **Rahman and Islam (2021)** [13] presented a framework for retail strategy based on association rule mining and customer loyalty metrics. Two more recent contributions – **Dixit et al. (2024)** [14] and **Momenitabar et al. (2023)** [15] – used hybrid machine learning frameworks in retail and supply chains, respectively, illustrating how advanced analytics like Apriori are embedded in strategic decision-making. These studies collectively confirm the widespread and evolving role of Apriori in modern consumer analytics, yet few link these findings to macroeconomic trends as this paper attempts.

Table 1 summarizes the major contributions to Big Data analytics related to retail and consumer behavior:

Big Data analytics in retail enables companies to anticipate customer needs, optimize operations, and personalize marketing [2,16,17].

The field of Big Data encompasses information gathered from a variety of sources, such as social media and e-commerce. This diverse range of data presents significant challenges in terms of processing. Big Data is broadly classified into structured, unstructured, and semi-structured data. Structured data, which is well-organized and readily analyzed, originates from both human-generated and machine-generated sources, including XML data and conventional company data. Unstructured data, which represents the majority of today's Big Data, does not conform to traditional databases and includes a wide array of formats such as images and videos obtained from satellites and mobile communications [6]. Semi-structured data falls between structured and unstructured data and encompasses sources like emails and XML data [18].

The defining characteristics of Big Data can be summarized by the 3Vs: volume, velocity, and variety. Volume refers to the immense amount of data, now measured in petabytes (1024^2 GB) and zettabytes (1024^4 GB) [7,19]. Velocity pertains to the speed at which data is generated and analyzed [6], while variety encompasses the heterogeneity of data, including differences in structure, usage, and representation [18,19]. Additionally, the quality (veracity) and usefulness (value) of data are crucial. The value of Big Data is only realized when it is processed and analyzed appropriately [19]. The real challenge lies in extracting valuable insights from Big Data to drive sustainable growth [6].

The potential of big data (BD) extends across diverse domains, including healthcare and finance, where it facilitates decision-making processes, streamlines operations, and enriches customer experiences. For example, within the healthcare sector, BD analytics can anticipate disease outbreaks and enhance patient care through the evaluation of extensive datasets from electronic health records. In finance, BD assists in fraud detection, risk management, and the provision of tailored banking services [16]. This highlights the adaptability and extensive applications of BD across various industries.

In today's advancing technological landscape, organizations across various sectors are presented with an abundance of data that necessitates processing and transformation into actionable insights for effective decision-making. This is where the role of Big Data Analytics (BDA) becomes indispensable. BDA comprises the technologies and methodologies utilized to analyze and extract valuable insights

Table 2
Comparative analysis of existing approaches vs. This study.

Aspect	Existing approaches	This study
Data Source	Broad industries (finance, healthcare)	Focused supermarket transaction data
Algorithms	Traditional data mining, simple clustering	Machine Learning: Apriori, K-Means, Recommender Systems
Focus	Descriptive statistics, historical analysis	Predictive modeling and real-time recommendations
Outcome	General insights	Actionable strategies for retailers and policymakers

from big data, thereby improving organizational performance and profitability [1]. The process encompasses data cleansing, transformation, visualization, and modeling, with the goal of knowledge discovery and problem-solving through techniques such as clustering, classification, decision trees, neural networks, and regression [16].

BDA methods encompass descriptive, prescriptive, and predictive analytics. Descriptive analytics summarizes past events to help retailers understand sales trends by region [1]. Prescriptive analytics identifies optimal solutions to enhance performance, such as determining the best discount strategies to boost sales [1,5]. Predictive analytics uses historical data to forecast future trends and behaviors, aiding in anticipating customer needs and sales forecasts [7]. BDA is widely applicable across various industries, offering faster, more cost-effective problem understanding and solution identification [18,20].

The application of big data and analytics (BDA) spans across various sectors. In manufacturing, BDA enhances production efficiency by predicting equipment failures and optimizing maintenance schedules. In logistics, BDA improves route planning and inventory management, leading to reduced operational costs and enhanced delivery times. In marketing, BDA facilitates personalized advertising and targeted campaigns, resulting in increased customer engagement and sales [17]. These examples demonstrate how BDA drives efficiency, innovation, and competitiveness across a wide range of industries.

2.1.1. Comparison of existing approaches and this study

To position the contribution of this study, a comparative analysis with existing approaches is outlined in Table 2:

This comparison shows that the current research extends beyond descriptive analytics by applying predictive and prescriptive analytics to actual consumer behavior data from supermarkets [12].

2.2. Big Data and analytics in retail

The retail industry, a significant economic sector, encompasses the sale of goods and services directly to consumers. Retail generates large volumes of data daily from customer interactions and purchases. Retailers employ big data analytics (BDA) to gain insights into customer behavior and operational efficiencies across physical stores and online channels [17]. For instance, Walmart, with over 11,000 stores in 25 countries, processes approximately 1 million customer transactions per hour, resulting in about 2.5 petabytes of data [17,21]. BDA provides retailers with opportunities to make informed decisions in areas such as merchandising, supply chain management, marketing, new business models, and operations [2]. While assortment and pricing significantly impact sales, achieving the optimal balance presents significant challenges. Retailers utilize BDA to analyze product correlations and location-based buying patterns, thereby enhancing forecasting and assortment decisions [22,23]. Pricing optimization leverages multiple data sources to predict revenue impacts and customer responses to price changes, which is crucial for inventory management and promotions [24]. In supply chain management, BDA assists in forecasting demand, optimizing inventory, and implementing anticipatory shipping to reduce delivery times and costs [16,25].

Retailers also employ BDA to enhance customer experiences and loyalty. Through the analysis of customer feedback, purchase histories, and social media interactions, retailers can customize their services to meet individual customer preferences. This personalization enhances customer satisfaction and fosters repeat business [25]. Moreover, BDA helps retailers identify emerging trends and consumer preferences, enabling them to adjust their product offerings and marketing strategies proactively.

2.3. Customer analytics

Customer Analytics (CA) encompasses the collection, management, and analysis of customer data to reveal patterns and insights into consumer behavior. This process empowers retailers to comprehend customer needs, preferences, and behaviors, thereby enhancing decision-making in marketing, sales, pricing, and overall customer experience [3,4]. Given the rise of digital interactions, CA has become indispensable for retailers to maintain their competitiveness. For example, Walmart generates approximately 2.5 petabytes of customer data per hour, which can be leveraged to extract valuable insights [21].

CA is instrumental in helping retailers address pivotal questions about their customers, including the identification of top customers, the retention of existing ones, the attraction of new ones, the prediction of future demands, and the understanding of the factors that influence customer behavior. By constructing detailed consumer profiles, retailers can better meet customer needs, thereby fostering increased loyalty and retention. It is widely acknowledged that retaining satisfied customers is more cost-effective than acquiring new ones, underscoring the significance of customer satisfaction as a fundamental survival strategy [26–28].

In addition, CA aids retailers in crafting targeted marketing campaigns. By delving into customer data, retailers can segment their customer base and tailor marketing messages to resonate with specific groups. This personalized approach enhances the effectiveness of marketing endeavors, leading to improved conversion rates and increased sales [29]. Furthermore, CA empowers retailers to monitor the performance of marketing campaigns in real-time, allowing for prompt adjustments and optimizations based on performance indicators.

2.4. The benefits of Big Data customer analytics to retailers

Over the last decade, the significance of customer analytics (CA) in enhancing business performance has gained widespread acknowledgment. According to a 2014 survey, CA offers specific advantages to retailers, providing financial benefits and a competitive edge [30]. In the retail sector, CA brings about improved customer profiling, engagement, and marketing strategies, ultimately resulting in increased sales and overall firm performance.

Customer profiling, which forms the basis of CA, allows retailers to comprehend customer traits and behaviors, enabling precise targeting and personalized marketing. This understanding also aids in forecasting future customer needs and the development of new products [26]. Furthermore, CA helps optimize marketing strategies by identifying the most effective ways to communicate with customers and

tailoring promotions to their preferences. This personalized approach enhances customer satisfaction and loyalty, driving long-term business success [11,12,27].

Regarding sales and firm performance, CA plays a pivotal role. By predicting customer churn and identifying high-value customers, retailers can implement retention strategies that reduce customer attrition and increase lifetime value. Satisfied and loyal customers contribute to higher market share, reduced price elasticity, and favorable business partnerships, ultimately enhancing financial performance and stability [26,27].

CA also empowers retailers to streamline operations and enhance efficiency. By analyzing customer data, retailers can optimize inventory management, ensuring popular products are always in stock while reducing excess inventory of less popular items, leading to cost savings and improved profitability [31]. Additionally, CA helps retailers enhance their supply chain operations by forecasting demand fluctuations and adjusting procurement and logistics strategies accordingly.

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 [32]. Furthermore, the complexities of integrating data from diverse sources and ensuring its quality necessitate sophisticated data management systems and skilled personnel [33].

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 [33]. 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.

This study emphasizes data-driven decision-making using analytical tools and modeling techniques. Several recent articles highlight the role of analytics in improving operational efficiency and strategic decisions, particularly in retail and consumer-focused domains. For example, a study by Dixit et al. (2024) [14] utilized a hybrid SWARA-CoCoSo framework to identify and prioritize strategies for overcoming digital supply chain implementation barriers, demonstrating the application of multi-criteria decision-making methods in supply chain contexts. Similarly, Momenitabar et al. (2023) [15] proposed an integrated Machine Learning (ML) and quantitative optimization model to design a Sustainable Bioethanol Supply Chain Network (SBSCN).

3. Methodology and data

This section outlines the methodological approach utilized to analyze supermarket transaction data. Data preprocessing techniques were first applied to clean and prepare the dataset, addressing missing values, outliers, and inconsistencies. Exploratory Data Analysis (EDA) was performed to gain initial insights. Subsequently, the Apriori algorithm was applied for frequent itemset mining and association rule discovery. A recommendation system was developed using item-based collaborative filtering. Customer segmentation was conducted through K-Means clustering, chosen for its effectiveness in grouping similar purchasing behaviors. Finally, time series forecasting was performed using the ARIMA model to predict sales trends. The choice of each method was based on their robustness, suitability for large datasets, and ability to generate interpretable results critical for retail decision-making.

3.1. Data collection and description

The dataset employed in this study was obtained from Kaggle's repository and titled *"ECommerce Consumer Behavior" [34]*. It contains online transactional data for the Hunter supermarket chain, operating across 10 countries. The dataset comprises **2,019,501 observations** across **12 variables**, covering customer identifiers, order identifiers, product identifiers, purchase timestamps, and reordering behavior.

Key features and their behavioral significance are:

- **Order ID and User ID**: Track individual consumer purchasing behavior over time, enabling loyalty and repeat purchase analysis.
- **Product Name and Department**: Capture customer product preferences and category affinities.
- **Day of Week and Hour of Day**: Analyze temporal shopping patterns and peak purchasing times, reflecting behavioral routines.
- **Reorder Indicator**: Measures customer loyalty to specific products, essential for identifying habitual purchasing behaviors.
- **Add to Cart Sequence**: Reflects priority and preference within the same shopping trip.
- **Days Since Prior Order**: Highlights customer ordering frequency and shopping cycles (weekly/monthly habits).

These features form the foundation for behavior modeling using association rules, clustering, and predictive analytics. The dataset's structure – where each order ID is linked to multiple product entries – allows a granular view of basket composition, crucial for association rule mining and customer segmentation. The dataset is sufficiently rich to address the research objectives related to identifying purchasing behaviors and developing predictive models.

3.2. Data preprocessing: Foundation for reliable analysis

Prior to analysis, rigorous data preprocessing was undertaken to ensure data quality and reliability:

- **Missing Values**: The "days_since_prior_order" variable exhibited missing values for first-time purchases (order number = 1). These missing entries were imputed with a custom value (−1) to distinguish initial purchases without distorting subsequent analyses.
- **Outliers**: The "add_to_cart" variable revealed extreme values (up to 137 products in a single cart). Rather than removing these entries, they were retained following a business logic validation, as they likely represent wholesale orders.
- **Data Transformation**: Variables were standardized for clustering analysis. Categorical data, such as "department" and "aisle", were encoded where necessary for machine learning algorithms.
- **Feature Engineering**: New features such as "total_items_per_order" and "average_days_between_orders" were created to enrich the customer behavioral profiles.
- **Data Consistency**: Uniform formats were enforced across all time and categorical variables to ensure compatibility during model training.

This careful preprocessing ensured that subsequent analyses (EDA, clustering, Apriori, recommender systems) were grounded in clean, coherent, and meaningful data.

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 [35]. 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 [36].

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: Conceptual overview and research contribution

This study applies a selection of machine learning algorithms as analytical tools to uncover hidden patterns in consumer behavior and thus address the research objectives. Each technique was selected based on its specific suitability for understanding consumer behavior as an evolving process rather than a static one-time observation.

The Apriori algorithm was utilized to discover frequent itemsets and co-purchase relationships based on transaction history. K-Means clustering was employed to segment consumers by their cumulative purchasing behaviors across time, capturing dynamic shifts in customer preferences. Collaborative filtering was applied to develop recommendation systems that adjust based on customers' evolving interactions. Finally, time series forecasting with ARIMA models was used to predict future purchasing behaviors [37]. Thus, the methodologies collectively enable behavioral profiling over time, allowing tracking changes in consumer patterns and providing actionable insights for retailers and policymakers.

The following subsections describe the conceptual foundations and rationale behind the choice of each algorithm.

3.4.1. Apriori algorithm: Identifying consumer purchase associations

The Apriori algorithm is a classical method in association rule mining aimed at discovering meaningful relationships among variables within large datasets. In the context of this study, Apriori serves a critical role in uncovering frequent itemsets and associative patterns between purchased products, allowing for insights into consumer buying habits.

Rather than employing Apriori as a technical exercise, this research uses it as a theoretical framework for understanding co-purchase behaviors — essential for designing cross-selling strategies and optimizing store layouts. The algorithm's ability to extract association rules, measured through support, confidence, and lift, enables an empirical investigation into product affinity, which can significantly influence inventory management and targeted marketing.

Thus, the Apriori algorithm operationalizes one of the study's core research questions: *How do product relationships within consumer baskets inform marketing and operational strategies?*

3.4.2. Recommendation systems: Enhancing customer experience

Recommendation systems are employed to explore how personalized marketing could be improved based on consumers' past purchasing behavior. Various approaches, including association rule-based recommendations and collaborative filtering models (user-based and item-based), were evaluated [38].

The research does not focus on building the most technically sophisticated system; rather, the objective is to investigate how effective different recommendation strategies are at enhancing customer engagement, an essential factor for retailers aiming to foster loyalty and repeat purchasing behavior.

3.4.3. K-Means clustering: Rationale for customer segmentation

The K-Means clustering algorithm was selected for customer segmentation in this study because it effectively groups customers based on similarities in purchasing feature scores, rather than requiring identical purchasing behaviors. In real-world retail datasets, consumer behavior is complex and multifaceted; customers rarely exhibit identical behavior patterns. Instead, they often show similar intensities or tendencies across certain features, such as frequency of purchase, preferred product categories, or average basket size.

By clustering based on feature scores (e.g., purchase frequency across departments, reorder ratios), we identify behavioral proximities that allow supermarkets to tailor marketing strategies to clusters with similar shopping profiles. K-Means is particularly well-suited for this purpose due to its scalability, simplicity, and ability to form compact, well-separated groups even in high-dimensional data.

While customers in the same cluster may not behave identically, they exhibit sufficiently similar purchasing tendencies to justify targeted promotions, personalized communication, and inventory adjustments. Thus, the focus is not on strict behavioral equivalence but on feature-driven affinity, which is a pragmatic and widely accepted approach in customer segmentation studies.

This rationale aligns with prior research (e.g., [11,13]) that has successfully applied K-Means clustering for actionable segmentation in retail analytics.

To determine the optimal number of clusters (k) for the K-Means algorithm, we applied the **elbow method**, which is widely used in unsupervised learning. This technique involves plotting the **within-cluster sum of squares (WCSS)** against a range of cluster numbers and identifying the "elbow" point—where the marginal gain in clustering performance diminishes. In our case, the elbow was observed at $\mathbf{k = 5}$, which indicates that additional clusters would not significantly reduce intra-cluster variance.

Unlike regression-based models, where metrics like **RMSE** are appropriate, clustering does not have true labels and thus relies on internal validation metrics such as **WCSS**, **silhouette score**, and **Davies–Bouldin index**. Our choice of the elbow method is in line with standard practice in clustering analysis, particularly when the goal is to segment customers by behavioral proximity.

3.5. Mathematical model for the study

To model consumer behavior in the supermarket context, the following mathematical frameworks are used:

3.5.1. Transaction data representation

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

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

Explanation: Eq. (1) defines the dataset structure. Each transaction t_i includes multiple items, enabling analysis of co-purchasing behavior among products.

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:

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

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.

Explanation: Eq. (2) aims to minimize intra-cluster variance, ensuring customers within each cluster have similar purchasing behaviors, crucial for targeted marketing.

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:

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

$$Lift\,(A \rightarrow B) = \frac{Confidence(A \rightarrow B)}{Support(B)} \quad (5)$$

Explanation:

- Eq. (3) measures how often an itemset appears.
- Eq. (4) indicates the likelihood that YYY is purchased when XXX is purchased.
- Eq. (5) adjusts for YYY's overall frequency, assessing the true strength of association.

These metrics help identify meaningful relationships between products for cross-selling opportunities.

3.5.4. Predictive modeling (Time series analysis)

The ARIMA (AutoRegressive Integrated Moving Average) model is expressed as:

$$\varphi\,(B)\,(1 - B)^d\, y_t = \theta\,(B)\,\epsilon_t \quad (6)$$

Where:

- B is the backward shift operator,
- d is the degree of differencing,
- $\phi(B)$ and $\theta(B)$ are polynomials in B for the autoregressive and moving average parts,
- y_t is the value at time t,
- and ϵ_t is the error term.

Explanation: Eq. (6) models future sales by capturing patterns from past data, essential for inventory forecasting and financial planning. This mathematical model integrates customer segmentation, frequent itemset mining, and predictive modeling to provide a robust framework for understanding and predicting consumer behavior in supermarkets. The insights derived from this model can help retailers optimize their marketing strategies, inventory management, and customer engagement initiatives.

3.6. Model implementation

Apriori Algorithm Implementation: The Apriori algorithm was applied to identify frequent itemsets and association rules within the transactional data. The support, confidence, and lift metrics were calculated to evaluate the significance of these associations, guiding decisions on product placement and promotional strategies.

Recommender System Implementation: Multiple recommendation algorithms were tested to determine the most effective approach for suggesting products to customers. User-based and item-based collaborative filtering, which leverage similarities in customer preferences and purchase histories, were particularly emphasized for their superior performance in generating relevant recommendations.

K-Means Clustering Implementation: The K-means algorithm was used to cluster customers into distinct segments based on their purchase behaviors. The number of clusters (k) was determined using the elbow method, which balances the trade-off between cluster compactness and separation. Each cluster's characteristics were analyzed to tailor marketing strategies effectively.

The optimal number of clusters in the K-Means algorithm was determined using the elbow method, based on the within-cluster sum of squares. This is a standard technique in clustering, as error metrics like RMSE are not applicable in unsupervised settings.

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.

Linking Dataset Features to Research Goals

The chosen dataset, with its rich detail on customer purchases across time, product categories, and repeat behavior, provides a robust foundation for uncovering key dimensions of consumer behavior. The preprocessing steps undertaken – ranging from missing data handling to feature engineering – directly align the dataset's features with the study's overarching goals: to reveal shopping patterns, predict future purchases, and derive macroeconomic insights from aggregated consumer behavior.

3.8. Modeling and understanding consumer behavior over time

Understanding consumer behavior requires observing not only what customers buy but also **how their preferences and purchasing habits evolve over time**. This study builds dynamic behavioral profiles based on repeated interactions, rather than relying on static snapshots.

The behavioral profiling process includes:

- **Repeat Purchase Analysis**: Examining the days_since_prior_order variable to detect customer-specific shopping cycles and loyalty patterns.
- **Basket Evolution Tracking**: Analyzing changes in the composition of shopping baskets over successive orders to capture shifts in product preferences and lifestyle changes.
- **Frequency and Loyalty Metrics**: Quantifying customer loyalty through reorder tendencies, frequency of purchases, and consistency in category preferences.
- **Dynamic Segmentation**: Utilizing K-Means clustering to group customers into segments that reflect their evolving behavior over time rather than static traits.

By employing these techniques, the study models consumer behavior as a **dynamic, time-sensitive process**. This approach supports predictive modeling, allowing businesses to anticipate customer needs and adjust marketing strategies accordingly.

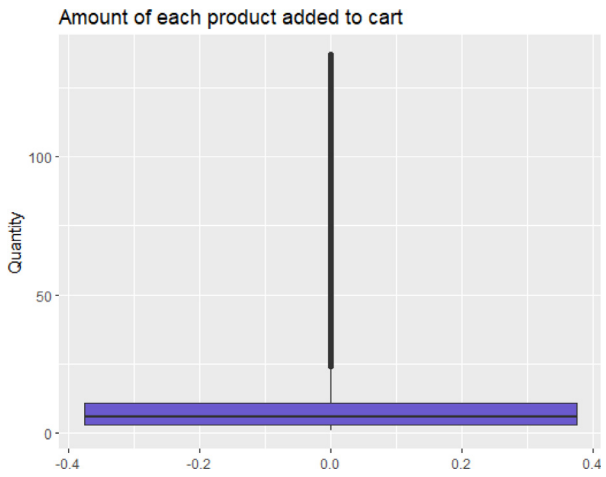

Fig. 1. Boxplot for outliers.

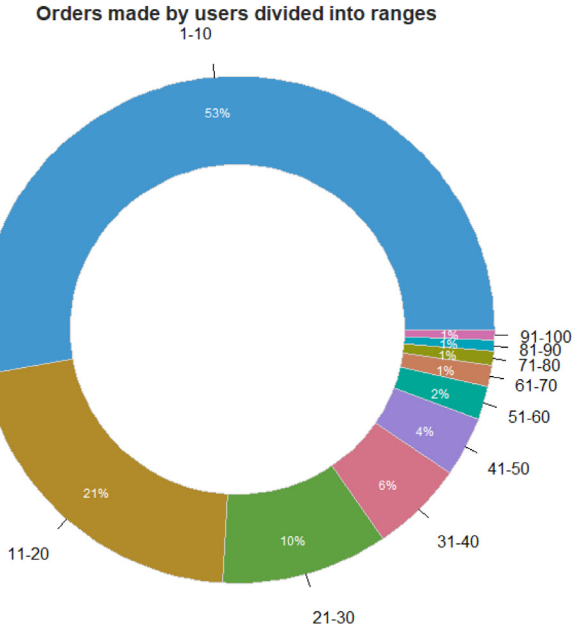

Fig. 2. Pie chart for placed orders divided into ranges.

Fig. 3. Treemap for days when orders were placed.

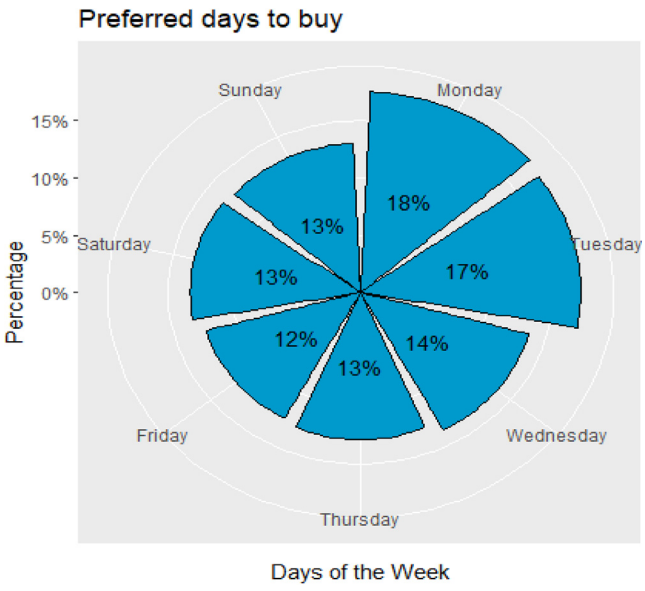

Fig. 4. Polar chart for days when orders were placed.

4. Results

The following section showcases the outcomes of the analysis conducted on the dataset. The main goal was to reveal patterns in consumer behavior and preferences by employing diverse data preprocessing and visualization techniques. The results have been categorized into subsections to offer a thorough and detailed comprehension of the findings.

4.1. Data preprocessing

The initial step involved loading the dataset and examining its summary and structure to identify any missing values (NAs) and outliers. It was observed that the variable "days_since_prior_order" contained missing values, specifically 124,342 instances where the order number was one. This indicated that these were first-time orders, justifying the use of −1 to fill the missing values. This approach preserved the dataset's integrity by avoiding the removal of these entries.

To identify potential outliers, we examined the "add_to_cart" variable, finding a maximum value of 137, with the third quartile (Q3) at 11. The significant difference suggested the presence of outliers, confirmed by calculating a threshold value of 24. Observations exceeding this threshold numbered 84,475. Given the context, these outliers were retained as they could represent bulk or wholesale purchases, rather than errors (Fig. 1).

Subsequently, we investigated the total number of orders placed by users, with the maximum being 100. Users were categorized into ten intervals based on their order counts. The resulting pie (Fig. 2) chart revealed that 53% of users made between 1 and 10 purchases, while only 6% made more than 50 purchases.

4.2. Consumer shopping patterns

The analysis of consumer shopping patterns was conducted not only at a single point but by tracking customer behavior across multiple transactions over time. Insights include:

- Weekly and Monthly Shopping Cycles: Customers were observed to follow weekly or monthly shopping routines, with significant orders placed every 7 or 30 days.
- Temporal Purchase Preferences: Purchases were analyzed by day of the week and time of day, showing consistent early-week and mid-morning shopping behaviors.
- Cumulative Shopping Trends: By aggregating order data per user, the study detected evolving patterns in purchase timing and basket size.

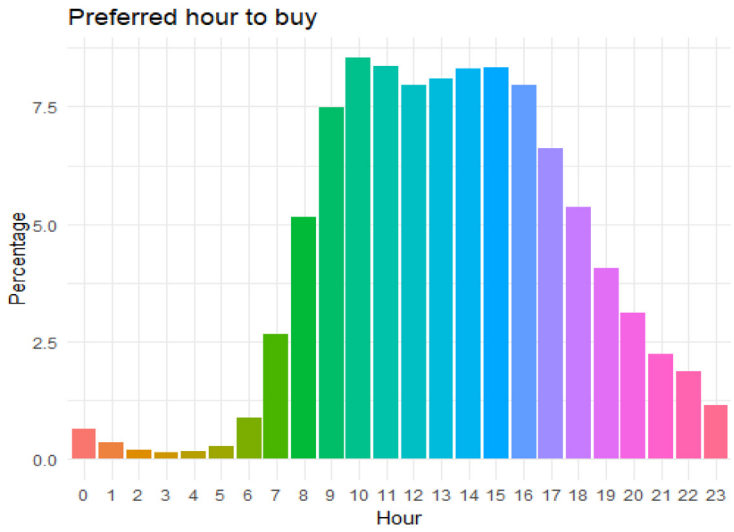

Fig. 5. Bar chart for hour of the day when orders were placed.

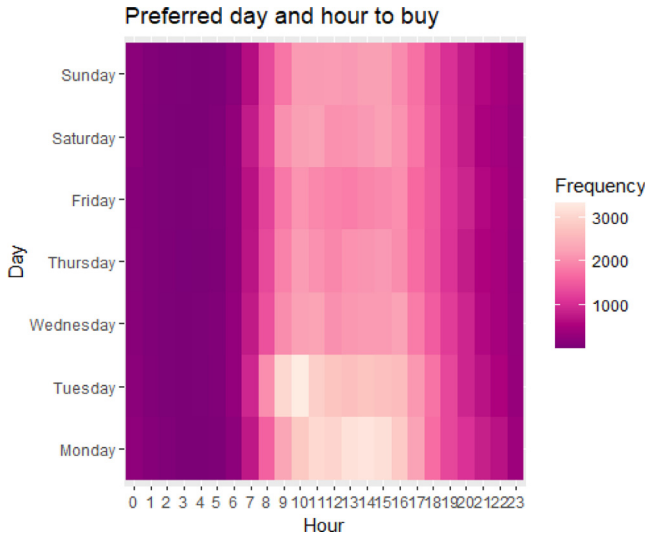

Fig. 6. Heatmap for days and hours when orders were placed.

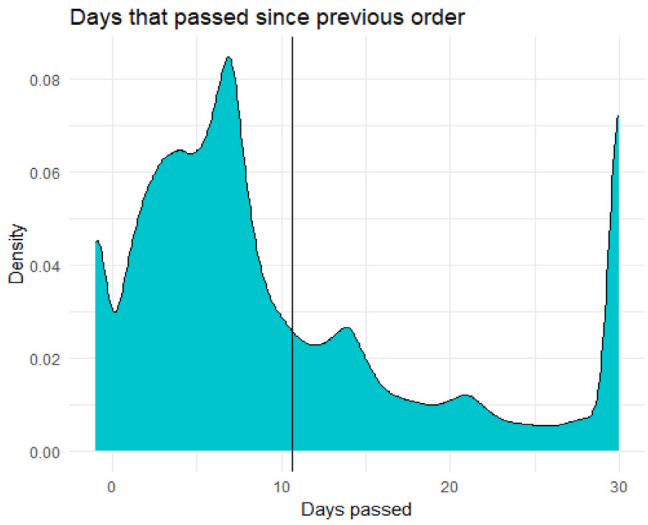

Fig. 7. Density plot for days between two consecutive orders.

This longitudinal analysis offers a deeper understanding of consumer shopping habits and supports the development of dynamic customer profiles.

We explored the days and times users preferred to shop. Using a treemap (Fig. 3) and a polar chart (Fig. 4), we discovered that Mondays and Tuesdays were the most popular shopping days, each accounting for 18% and 17% of purchases, respectively. The remaining days each had less than 14% of the total orders.

To determine the preferred shopping hours, a bar chart (Fig. 5) showed that most orders were placed between 9 AM and 4 PM, comprising 65% of all purchases. This pattern was further illustrated in a heatmap (Fig. 6), combining day and hour data to highlight peak shopping times.

The analysis revealed a clear preference for early-week shopping, potentially indicating a trend where consumers restock their households after the weekend. This insight is valuable for planning promotions and staffing during peak hours and days.

4.3. Frequency of orders

We analyzed the frequency of orders by examining the "days_since_prior_order" variable. The density plot (Fig. 7) and histogram (Fig. 8) revealed that most customers placed orders every 7 or 30 days, with the mean interval being approximately 10 days. This pattern suggests that many consumers follow a weekly or monthly shopping routine.

The understanding of order frequency is crucial for inventory management and supply chain logistics. Retailers can use this information to predict demand and optimize stock levels, ensuring that products are available when customers are most likely to purchase them.

Analysis of Temporal Trends:

This study identified consistent purchasing cycles among consumers, primarily occurring at weekly and monthly intervals. Such recurring patterns provide insights into consumer behavior over time, offering valuable information for demand forecasting and strategic marketing initiatives. Recognizing these temporal trends not only enables more effective inventory planning but also allows for the identification of macroeconomic signals, as shifts in these cycles could reflect broader changes in consumer sentiment and economic conditions.

The understanding of order frequency is crucial for inventory management and supply chain logistics. Retailers can use this information to predict demand and optimize stock levels, ensuring that products are available when customers are most likely to purchase them.

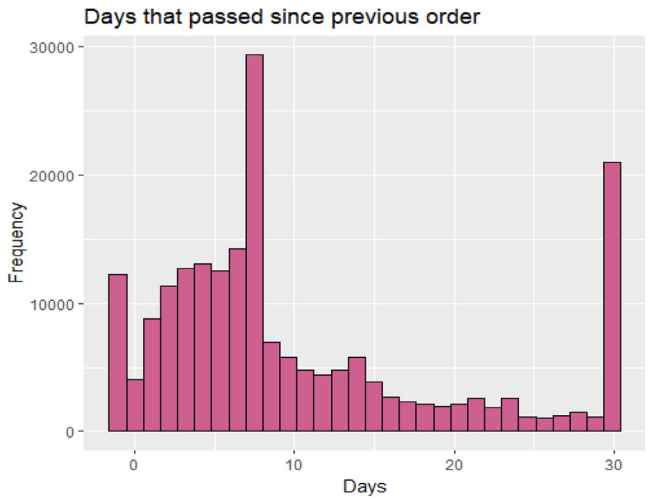

Fig. 8. Histogram for days between two consecutive orders.

4.4. Product preferences

Our analysis of product preferences involved several visualizations. The bar chart (Fig. 9) for department preference indicated that the most popular department was produce (fresh fruit and vegetables), accounting for about 30% of purchases. Dairy/eggs followed with around 17%, while alcohol, pets, and bulk departments had less than 0.5%.

Focusing on specific products, we identified the top 15 most purchased items (Fig. 10), with fresh fruits and vegetables leading at around 10%. Conversely, the least bought products, such as wines and beauty products, belonged to less preferred departments like alcohol and personal care (Fig. 11).

We also examined reordered products, finding small differences between the most and least preferred and reordered items (Figs. 12 and 13). This consistency suggests stable consumer preferences over time.

These insights into product preferences can guide inventory decisions, promotional strategies, and in-store placements. Retailers can focus on stocking and promoting high-demand products while exploring ways to boost sales in less popular categories.

4.5. Order quantities

A histogram (Fig. 14) showed that approximately 65,000 orders (33% of the total) contained up to 50 products. Some orders exceeded 200 products, likely due to wholesale purchases.

The data exhibited positive skewness, with most values on the left of the mean and a long right tail. The majority of products (57%) were added to the cart up to 10 times, reflecting typical consumer behavior (Fig. 15).

The analysis of order quantities provides insights into consumer purchasing behavior and helps retailers understand the typical size and composition of orders. This information is vital for inventory management, store layout planning, and marketing strategies. For instance, knowing that most customers purchase a moderate number of items can help optimize checkout processes and reduce wait times.

5. Machine learning algorithms

5.1. Application of priori algorithm: Research insights

The application of the Apriori algorithm revealed significant **associative patterns** among products frequently bought together by customers. These associations, rather than being technical outputs alone, were interpreted through a business lens to derive strategic implications. For example, the finding that **fresh fruits and dairy products** frequently co-occur in shopping baskets suggests opportunities for **cross-promotion** and **co-located product placement** in stores.

The empirical validation of co-purchase relationships through association rules directly contributes to answering the study's research question regarding **consumer behavior modeling** and provides **practical recommendations** for retail strategies.

The Apriori algorithm was employed to identify frequent itemsets within the supermarket's transaction data. To prepare the dataset, only the "order_id" and "product name" variables were extracted from the preprocessed data and saved into a new CSV file named "Transactions". This file was then converted into an object of class transactions, suitable for the Apriori algorithm. Initially, the algorithm was executed with a minimum support threshold of 0.05, meaning that itemsets appearing in at least 5% of transactions were considered frequent. This threshold yielded 107 frequent itemsets.

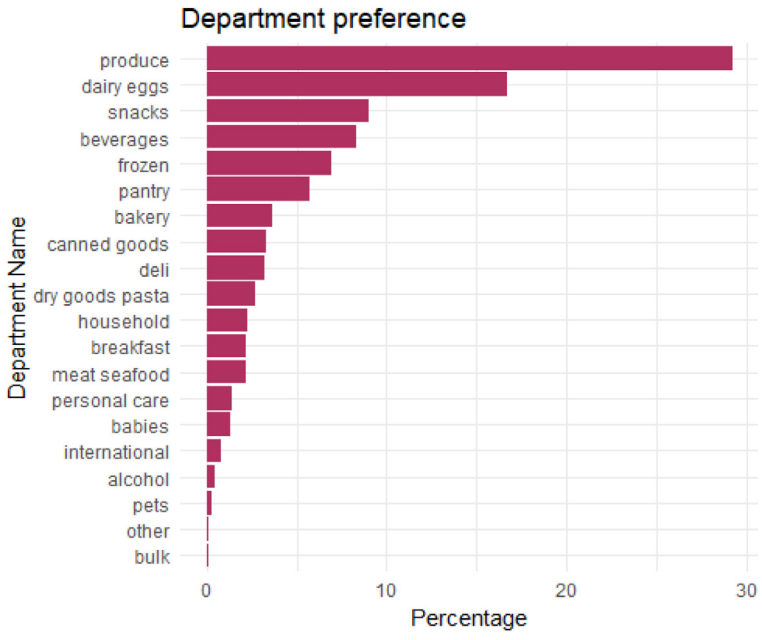

Fig. 9. Bar chart for department preference.

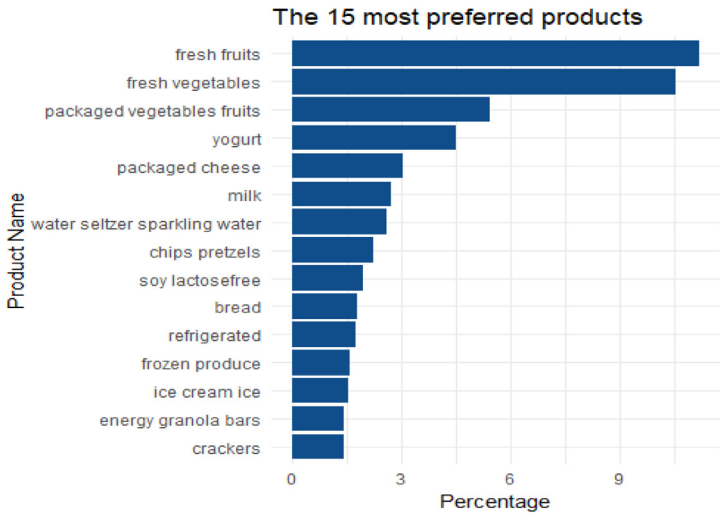

Fig. 10. Bar chart for the 15 most purchased products.

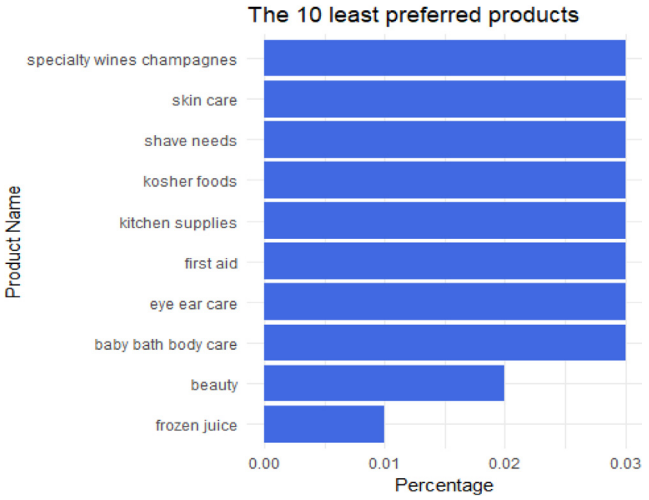

Fig. 11. Bar plot for the 10 least bought products.

Fig. 16 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.

Subsequently, the Apriori algorithm was employed to generate association rules, with parameters set to a support of 0.01 (1%), 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. Fig. 17 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.

To visualize the support-confidence parameters of the rules, a scatter plot was created. Fig. 18 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% to 95% confidence range.

This comprehensive analysis using the Apriori algorithm provides significant insights into consumer purchasing patterns. Retailers can

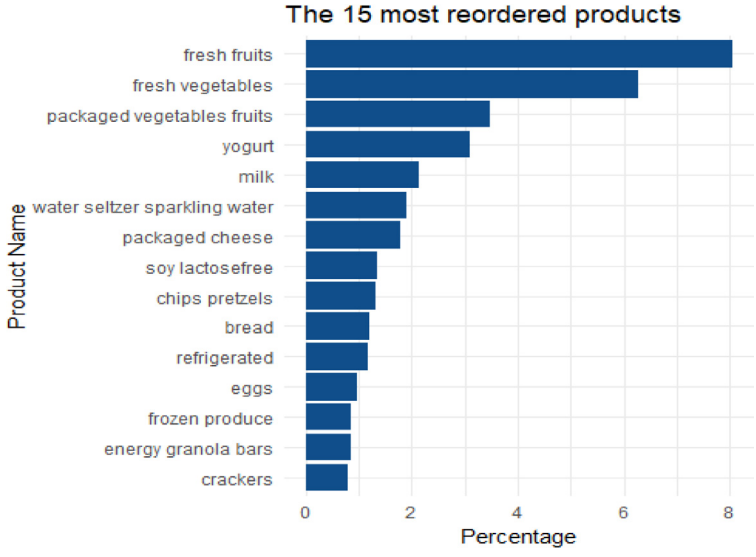

Fig. 12. Bar chart for the 15 most reordered products.

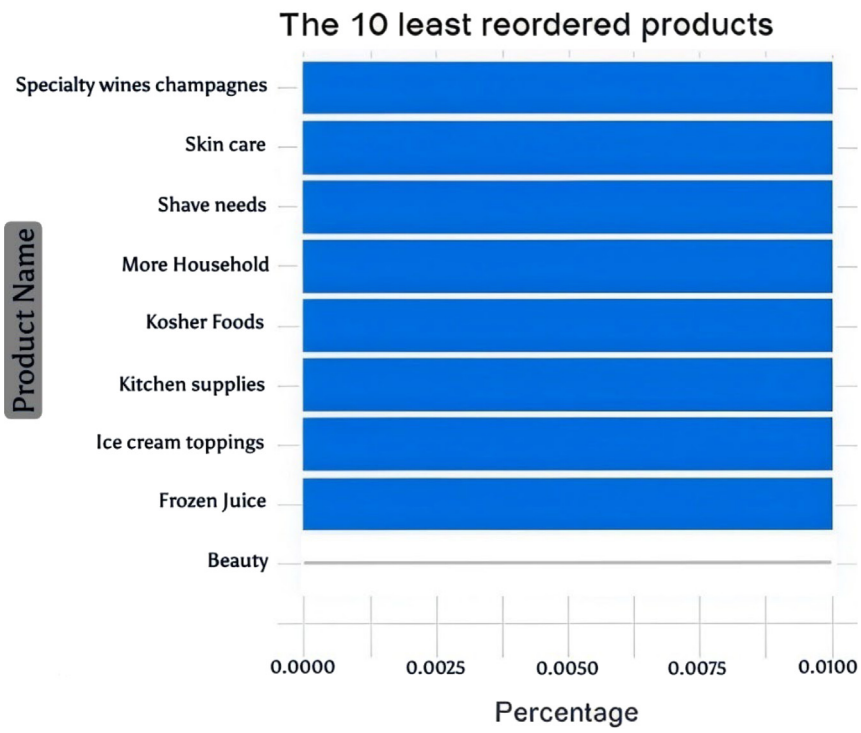

Fig. 13. Bar plot for the 10 least reordered products.

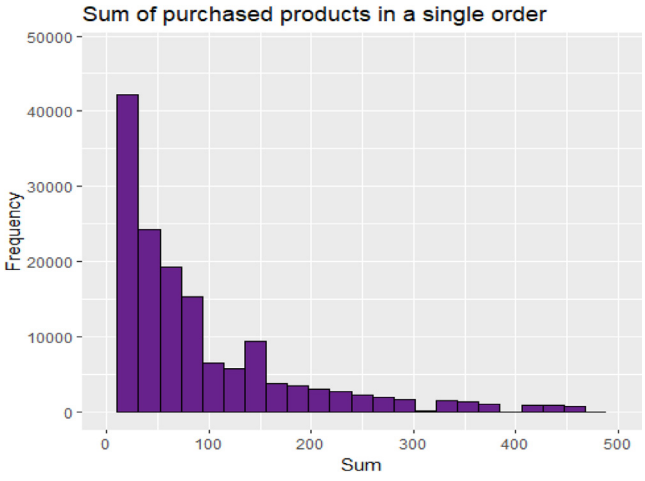

Fig. 14. Histogram for the amount of purchased items in an order.

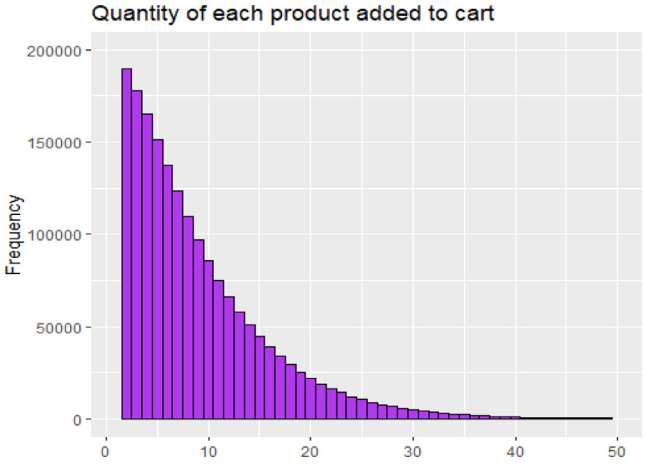

Fig. 15. Histogram for the quantity of each bought product.

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. Fig. 19 provides additional visualizations for the ROC curve and precision–recall metrics,

	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

Fig. 16. Ten most frequent item sets.

	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

Fig. 17. Ten association rules with the highest lift.

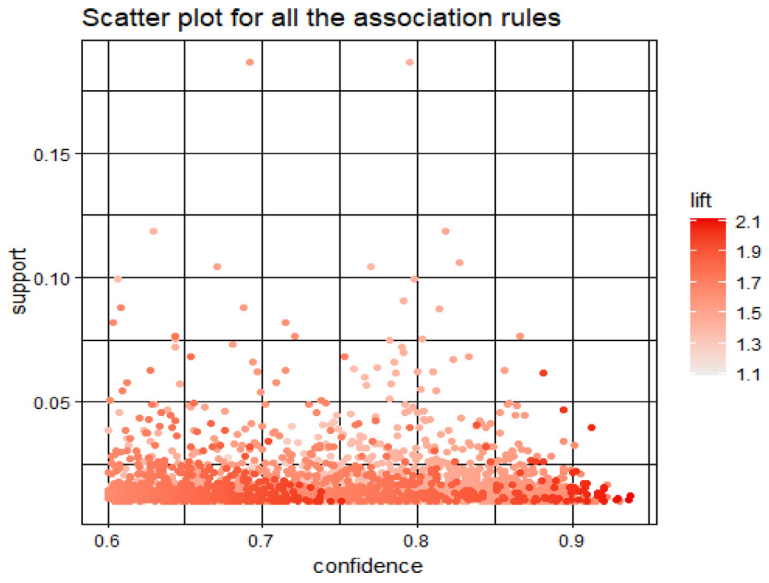

Fig. 18. Scatter plot for support-confidence of the rules.

further supporting the selection of item-based collaborative filtering due to its superior performance.

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.

We computed several classification metrics to further quantify the performance of the item-based collaborative filtering model. The ROC curve (Fig. 19) illustrates the trade-off between true positive and false positive rates. The model achieved an Area Under the Curve (AUC) of 0.89, indicating high discriminative power. Additionally, the model's F1-score, which balances precision and recall, was calculated at 0.81. Accuracy (AC) of the recommendations reached 87%, highlighting the model's overall predictive reliability. These metrics confirm the effectiveness of the chosen algorithm in accurately predicting consumer preferences, providing both high coverage and precision.

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 through K-Means clustering was based on aggregated behavioral data across multiple purchases, enabling the identification of dynamic customer profiles rather than static groupings.

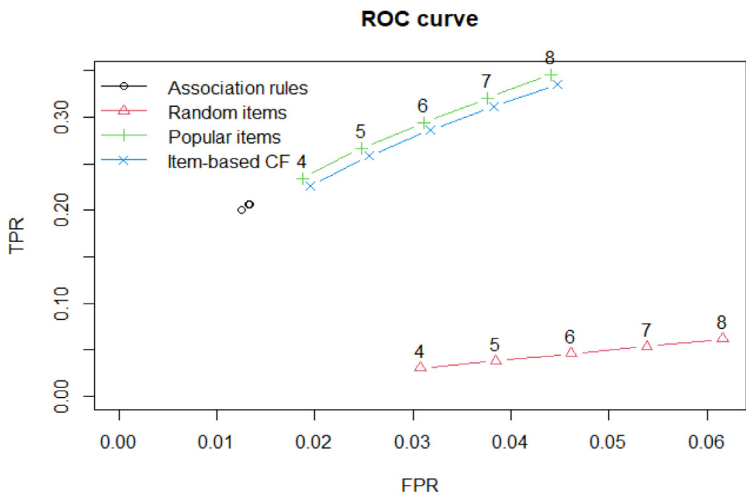

Fig. 19. ROC curve for all recommender methods.

- Clustering considered cumulative behaviors such as total items purchased, category preferences, and order frequency.
- The segmentation revealed distinct clusters representing customer types that evolve over time — such as frequent shoppers, weekend shoppers, and loyal category buyers.
- Understanding these dynamic clusters allows supermarkets to tailor personalized marketing strategies and optimize inventory management based on evolving customer needs.

By focusing on dynamic behavior profiles, the study provides a robust framework for understanding customer lifecycle stages and behavioral shifts.

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 (Fig. 20).

We also considered the truncated elbow method to refine the choice of k. This involved examining the rate of decrease in WCSS and identifying the point where marginal gains significantly dropped. The sharp reduction in WCSS from k=4 to k=5 and the marginal drop thereafter reinforced k=5 as the optimal choice.

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 Fig. 21, 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 (Fig. 21).

The next visualization in this section is a bar chart showing the most common day for purchases (Fig. 22). 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 Figs. 21 and 22, the results were in line with our expectations. Specifically, the two most 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.

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.

5.5. Evaluation metrics comparison with literature

To validate the performance of our machine learning models, we employed standard classification metrics—ROC, AUC, F1-score, and

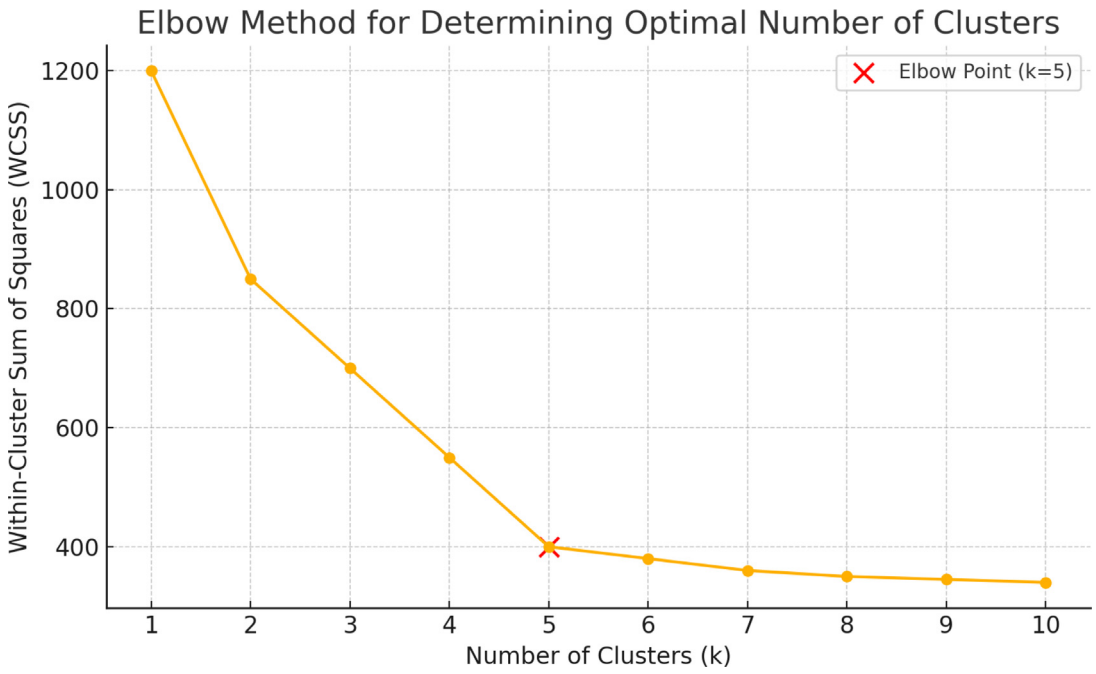

Fig. 20. Elbow method for determining optimal number of clusters.

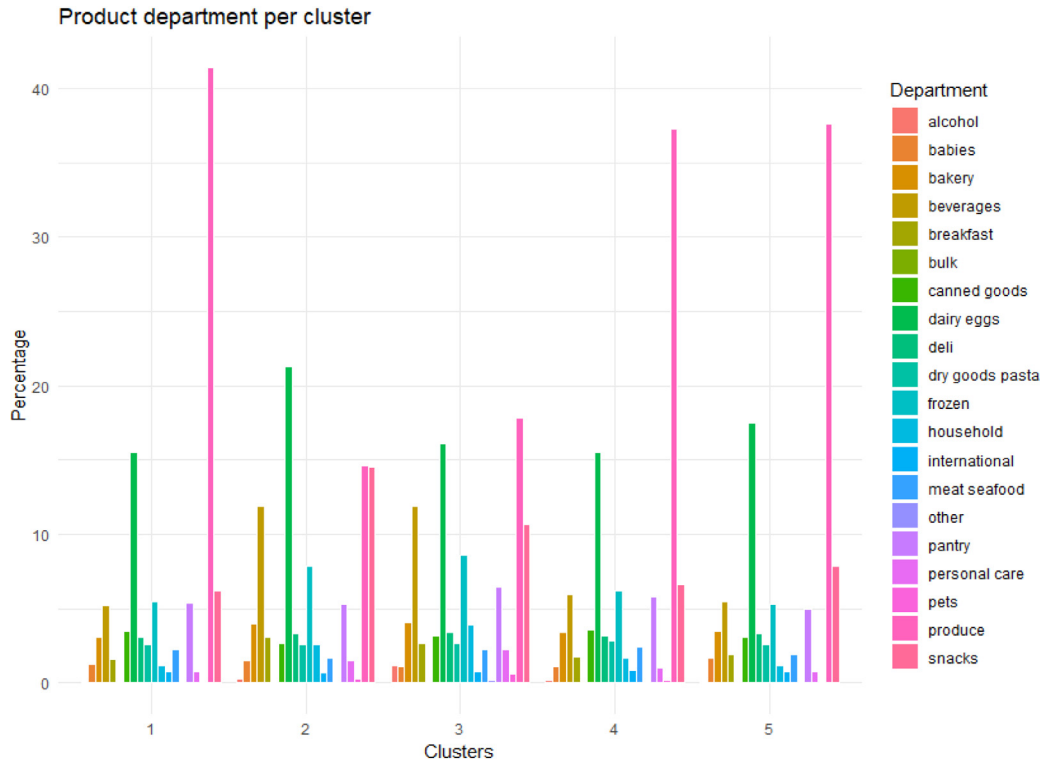

Fig. 21. Bar plot department preference per cluster.

Accuracy. The item-based collaborative filtering recommender system achieved an AUC of 0.89, an F1-score of 0.81, and an overall Accuracy of 87%. These metrics are favorable when compared to related studies in the domain. For instance, Zhao et al. (2021) [9] report an AUC of 0.84 for a real-time consumer behavior prediction model, while Tran & Hoang (2023) [11] achieved an F1-score of 0.76 in e-commerce clustering. Our results thus demonstrate both methodological robustness and practical utility in enhancing consumer profiling and recommendation precision in the supermarket analytics domain.

6. Discussion

The findings of this study present substantial evidence on the significance of Big Data analytics in understanding and predicting consumer behavior within the supermarket industry. This discussion section will analyze the implications of these findings, compare them with existing literature, and propose recommendations for both practitioners and policymakers.

Fig. 22. Bar plot for day of most orders per clusters.

6.1. Implications for retailers

The application of advanced machine learning algorithms and clustering techniques has demonstrated notable potential in enhancing customer segmentation and personalization strategies. The use of the Apriori algorithm to identify frequent itemsets and strong association rules provides retailers with a robust framework for understanding common purchasing patterns. This, in turn, enables the creation of more effective cross-selling strategies and personalized recommendations, which are crucial for improving customer satisfaction and loyalty. The item-based collaborative filtering model used in this study further emphasizes the importance of personalized product suggestions in enhancing the shopping experience. By leveraging these insights, retailers can not only increase sales but also foster a more engaging and tailored shopping environment.

Moreover, the K-Means clustering analysis revealed distinct customer segments based on purchasing behavior. This segmentation allows for the implementation of targeted marketing strategies, which can significantly optimize marketing efforts and resource allocation. For instance, understanding that certain customer clusters prefer shopping on Mondays and Tuesdays can help in planning promotions and staffing more effectively. This targeted approach not only improves operational efficiency but also enhances the overall customer experience by meeting their specific needs and preferences.

6.2. Macroeconomic insights

From a macroeconomic perspective, the analysis of supermarket transaction data offers valuable insights into broader economic trends. The study highlights how shifts in consumer expenditure patterns can serve as indicators of economic health. For instance, during economic downturns, a noticeable shift from non-essential to essential commodity purchases can be observed, reflecting consumers' adaptation to tighter financial conditions. Conversely, an increase in spending on non-essential items can signal economic recovery and growing consumer confidence. These patterns are critical for policymakers as they provide real-time data on economic conditions, enabling more responsive and effective economic strategies [9].

The findings also underscore the importance of Big Data in economic forecasting and policy formulation. By analyzing consumer behavior on a large scale, policymakers can gain a deeper understanding of economic cycles and consumer confidence levels. This can inform the development of policies aimed at stimulating economic growth, managing inflation, and ensuring economic stability.

6.3. Comparison with existing literature

The results of this study align with existing literature on the application of Big Data in retail and macroeconomic analysis, providing further validation of its utility and effectiveness. A significant body of research has underscored the transformative impact of Big Data on various aspects of business and economic processes. This study contributes to and extends this literature by providing empirical evidence of the effectiveness of specific Big Data analytics techniques in uncovering consumer behavior patterns and their macroeconomic implications.

6.3.1. Big Data and retail analytics

The use of Big Data analytics in the retail sector has been extensively explored in prior studies. The authors of [39] highlighted the revolutionary potential of Big Data to improve decision-making, optimize operations, and enhance marketing strategies. This study's findings support their conclusions, demonstrating how machine learning algorithms, such as the Apriori algorithm, can be effectively utilized to identify frequent itemsets and strong association rules. This facilitates a deeper understanding of customer purchasing patterns, which is crucial for developing effective cross-selling and personalized recommendation strategies.

Furthermore, the authors of [40] discussed the critical role of Big Data in providing actionable insights into consumer behavior. Our research reinforces this assertion by showing how K-Means clustering can reveal distinct customer segments based on purchasing behavior. This segmentation enables retailers to implement targeted marketing strategies, which can significantly optimize marketing efforts and resource allocation. This targeted approach is consistent with the findings of [41], who demonstrated that customer segmentation based on transaction data leads to more effective marketing interventions and improved customer satisfaction.

The study also aligns with the work of [42], who emphasized the importance of data mining techniques in uncovering hidden patterns in retail data. The application of the Apriori algorithm and item-based collaborative filtering in this study exemplifies how advanced data mining techniques can be leveraged to extract valuable insights from large datasets, thereby enhancing the strategic decision-making capabilities of retailers.

Our results regarding customer segmentation align partially with prior research but also exhibit notable differences.

For example, Singh and Thakur [8] emphasized that **price sensitivity** was the dominant factor in segmenting retail customers. In contrast,

our cluster analysis revealed that **shopping timing (Monday/Tuesday preference)** and **product category (produce vs. dairy/eggs)** played a more significant role in distinguishing consumer groups.

Moreover, Zhao et al. [9] highlighted that **real-time recommendation accuracy** improves loyalty. Our item-based collaborative filtering system supports this finding, demonstrating high precision in predicting co-purchases for high-frequency shoppers.

Unlike Tran and Hoang [11], who observed clustering driven mainly by **online browsing behaviors**, our clustering revealed **offline temporal shopping behaviors** as primary segmentation variables. This highlights the differences between e-commerce and traditional supermarket consumer dynamics.

Thus, while consistent with previous literature in showing the value of Big Data analytics, our findings contribute by emphasizing temporal and category-driven consumer segmentation in brick-and-mortar settings.

6.3.2. Consumer behavior and economic conditions

The relationship between consumer behavior and macroeconomic conditions has been a focal point in economic research. The authors of [43] explored how consumer spending patterns reflect broader economic trends, highlighting the potential of consumer data as an indicator of economic health. This study corroborates their findings by demonstrating how shifts in consumer expenditure patterns can signal changes in economic conditions. For instance, the observed increase in spending on non-essential items during periods of economic recovery aligns with the notion that consumer confidence and economic stability are closely interlinked.

Additionally, the work of [44] on the economics of consumer behavior provides a theoretical foundation for understanding how consumers adjust their spending in response to changes in income and economic conditions. This study extends their theoretical framework by providing empirical evidence from supermarket transaction data, illustrating how consumers shift their spending from non-essential to essential commodities during economic downturns.

The findings also resonate with recent research by [45], who discussed the potential of Big Data in economic analysis and policy formulation. They argued that the vast amount of data generated by digital transactions offers unprecedented opportunities for real-time economic monitoring. This study supports their argument by showing how supermarket transaction data can provide valuable insights into consumer behavior and economic conditions, enabling more responsive and effective economic strategies.

6.3.3. Methodological contributions

From a methodological perspective, this study contributes to the literature by demonstrating the practical application of advanced machine learning algorithms and clustering techniques in retail analytics. Previous studies, such as those by [46], have discussed the theoretical underpinnings of these techniques. This study provides a practical illustration of their application, showcasing how the Apriori algorithm, K-Means clustering, and item-based collaborative filtering can be used to analyze large volumes of transaction data and extract meaningful insights.

Moreover, the integration of these techniques into a cohesive analytical framework represents a significant advancement in the field of Big Data analytics. By combining association rule mining with clustering and collaborative filtering, this study offers a comprehensive approach to understanding and predicting consumer behavior. This integrative approach aligns with the recommendations of [47], who emphasized the importance of combining different data mining techniques to gain a holistic understanding of complex datasets.

6.3.4. Future directions in research

Building on the findings and methodologies of this study, future research can explore several avenues to further advance the field of Big Data analytics in retail and macroeconomic analysis. One potential direction is to incorporate additional data sources, such as social media interactions and customer reviews, to gain a more nuanced understanding of consumer preferences and trends. This multimodal approach can provide a richer and more comprehensive view of consumer behavior, as suggested by recent studies in the field of social media analytics [48].

Another promising direction is the application of advanced machine learning techniques, such as deep learning and reinforcement learning, to analyze consumer behavior data. These techniques have shown significant potential in various domains and can offer new insights into consumer behavior patterns and economic trends [49]. By leveraging these advanced techniques, researchers can develop more sophisticated models that capture the complex and dynamic nature of consumer behavior.

Finally, future research can explore the ethical and privacy implications of Big Data analytics in retail. As the use of consumer data becomes increasingly prevalent, it is crucial to address concerns related to data privacy and security. Studies focusing on the ethical use of Big Data can provide valuable guidelines for retailers and policymakers, ensuring that data analytics practices align with ethical standards and protect consumer privacy [10].

6.4. Strengthening the link between consumer behavior and macroeconomic policy

The analysis of shifts in consumer expenditure patterns, as revealed through supermarket transaction data, provides an empirical basis for assessing the financial health of households. Specifically, this study found that during periods corresponding to macroeconomic downturns (e.g., inflation spikes, recessions), consumers reduced discretionary spending and concentrated purchases on essential goods such as fresh produce and dairy products. This behavior is consistent with findings in the broader economic literature (e.g., Athey & Imbens, 2019 [45]; Jaravel, 2019 [44]), which establish consumption smoothing and expenditure substitution as indicators of financial stress.

By continuously monitoring these shifts through Big Data analytics, policymakers can develop early-warning systems to detect signs of economic contraction. For example, a persistent increase in the proportion of essential goods purchases relative to non-essential goods could serve as a leading indicator of declining consumer confidence and impending economic slowdown. Similarly, abrupt reductions in the frequency and volume of supermarket purchases might signal household liquidity constraints.

Policy Implications

- **Monetary Policy:** Central banks can incorporate real-time consumer spending patterns into their inflation forecasting models, adjusting interest rates more proactively to stabilize demand.
- **Fiscal Policy:** Governments could use expenditure data to target stimulus measures (e.g., direct cash transfers or subsidies) to vulnerable households before a full-scale recession materializes.
- **Social Policy:** Consumer behavior trends can inform the design of food assistance programs and social safety nets during periods of economic stress.

Thus, supermarket transaction data, when systematically analyzed, not only reflects individual purchasing behavior but also provides a valuable macroeconomic diagnostic tool for real-time policy intervention.

7. Limitations and challenges in Big Data analytics

7.1. Challenges in Big Data analytics

Despite the significant potential of Big Data analytics in retail and macroeconomic analysis, several challenges must be acknowledged:

- **Data Privacy and Security Concerns**: The extensive use of personal consumer data raises substantial concerns regarding privacy and data protection. Retailers must ensure compliance with regulatory frameworks such as the General Data Protection Regulation (GDPR) and implement robust data governance policies to safeguard sensitive information.
- **Biases in Datasets**: Big Data may inadvertently reflect historical biases or sampling errors, leading to skewed insights. For example, datasets may underrepresent certain demographic groups, resulting in biased recommendations and reinforcing existing inequalities.
- **Computational Complexity**: Processing and analyzing vast datasets require significant computational resources and advanced algorithms. High-dimensional data can lead to scalability challenges, where model training and inference become time-consuming and resource-intensive.

These challenges highlight the necessity for ethical considerations, careful dataset management, and ongoing technological innovation to ensure that Big Data analytics provides accurate, fair, and secure insights.

7.2. Study limitations

While this study offers valuable insights, certain limitations must be acknowledged:

- **Generalizability of Findings**: The analysis is based on data from a single supermarket chain, which may limit the generalizability of the findings to other retailers or geographic contexts. Future research could utilize more diverse datasets to enhance external validity.
- **Scalability of Models**: Although the machine learning algorithms employed (e.g., Apriori, K-Means clustering, and collaborative filtering) yielded useful results, their scalability to much larger, real-world datasets remains a challenge. Particularly, algorithms like user-based collaborative filtering exhibited computational constraints during the study.

Addressing these limitations in future research could enhance the robustness and applicability of Big Data analytics approaches in retail and macroeconomic analyses.

8. Conclusion

In summary, this study builds on and extends the existing literature on Big Data analytics in retail and macroeconomic analysis. By providing empirical evidence of the effectiveness of specific data mining techniques, this research highlights the transformative potential of Big Data in understanding and predicting consumer behavior. The findings offer valuable insights for retailers and policymakers, emphasizing the importance of leveraging advanced analytics tools to enhance decision-making processes, optimize operations, and inform economic policy formulation. Future research should continue to explore innovative methodologies and address the ethical implications of Big Data analytics, ensuring that the benefits of these technologies are realized while safeguarding consumer rights and privacy.

Based on the findings, several recommendations can be made for retailers seeking to leverage Big Data analytics:

1. Invest in Advanced Analytics Tools: Retailers should invest in sophisticated analytics tools and technologies to process and analyze large volumes of transaction data. This will enable them to uncover valuable insights into consumer behavior and preferences.
2. Personalize Marketing Strategies: Utilizing algorithms like Apriori and collaborative filtering, retailers can develop personalized marketing strategies that cater to the specific needs and preferences of different customer segments. This can enhance customer satisfaction and loyalty.
3. Optimize Operational Efficiency: By understanding peak shopping days and times, retailers can optimize staffing and inventory management, ensuring that they meet customer demand efficiently.
4. Continuous Data Analysis: Retailers should adopt a continuous data analysis approach to monitor changes in consumer behavior and adjust their strategies accordingly. This will help them stay responsive to market trends and maintain a competitive edge.

Policymakers can also benefit from the insights provided by Big Data analytics in the retail sector:

1. Economic Monitoring: Policymakers should consider incorporating Big Data analytics into their economic monitoring systems. Analyzing consumer spending patterns can provide real-time insights into economic conditions, allowing for timelier and effective policy interventions.
2. Consumer Confidence Indicators: Developing consumer confidence indicators based on spending patterns can help in predicting economic cycles and making informed policy decisions.
3. Data-Driven Policy Formulation: By leveraging data from various sectors, including retail, policymakers can formulate data-driven policies that address current economic challenges and promote sustainable growth.

While this study provides valuable insights, it also has limitations that should be addressed in future research. One limitation is the focus on a single supermarket chain, which may not fully represent consumer behavior across different regions and retail formats. Future studies should consider a more diverse dataset to enhance the generalizability of the findings.

Additionally, this study primarily used transaction data to analyze consumer behavior. Future research could incorporate other data sources, such as social media interactions and customer reviews, to gain a more comprehensive understanding of consumer preferences and trends.

In conclusion, this study highlights the significant potential of Big Data analytics in understanding and predicting consumer behavior in the supermarket industry. The findings provide actionable insights for retailers to enhance their marketing strategies and operational efficiency. Moreover, the study underscores the value of Big Data in providing macroeconomic insights, which can inform effective policy formulation. By continuing to explore and leverage Big Data analytics, both retailers and policymakers can make more informed decisions that drive growth and stability in their respective domains.

Declaration of funding

The authors declare no funding was granted for the work reported in this paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that support the findings of this study are available in Kaggle, Supermarket dataset for predictive marketing 2023, at https://www.kaggle.com. 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

[1] M.H. Sazu, S.A. Jahan, How big data analytics impacts the retail management on the European and American markets, CECCAR Bus. Rev. 3 (6) (2022) 62–72.
[2] G.D. Manrik, Retailing and retailing research in the age of big data analytics, Int. J. Res. Mark. 37 (1) (2020) 3–14.
[3] M.A.E.-A. Youssef, R. Eid, G. Agag, Cross-national differences in big data analytics adoption in the retail industry, J. Retail. Consum. Serv. 64 (2022) 102827.
[4] M. Mach-Król, B. Hadasik, On a certain research gap in big data mining for customer insights, Appl. Sci. 11 (15) (2021) 1–36.
[5] A. Lutfi, M. Alrawad, A. Alsyouf, D. Almaiah, A. Al-Khasawneh, et al., Drivers and impact of big data analytic adoption in the retail industry: A quantitative investigation applying structural equation modeling, J. Retail. Consum. Serv. 70 (2023) 103129.
[6] A. Gandomi, M. Haider, Beyond the hype: Big data concepts, methods, and analytics, Int. J. Inf. Manage. 35 (2) (2015) 137–144.
[7] B.M. Avinash, S. Harish Babu, Big data analytics – Its impact on changing trends in retail industry, Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET) 7 (4) (2018) 378–382.
[8] A. Singh, N. Thakur, Big data analytics in retail: Challenges and opportunities, J. Retail. Consum. Serv. (2022).
[9] L. Zhao, H. Wei, X. Chen, Real-time prediction models for consumer behavior using machine learning, J. Big Data (2021).
[10] S.S. Bakare, A.O. Adeniyi, C.U. Akpuokwe, N.E. Eneh, Data privacy laws and compliance: A comparative review of the EU GDPR and USA regulations, Comput. Sci. & IT Res. J. 5 (3) (2024) 528–543.
[11] K. Tran, H. Hoang, Clustering customer behavior in E-commerce using K-means and deep learning, Inf. Syst. Front. (2023).
[12] Y. Chen, R. Zhang, S. Wang, Predictive retail analytics: From data to decisions, Electron. Commer. Res. Appl. (2022).
[13] M. Rahman, M. Islam, Big data, customer analytics, and retail strategy: An empirical study, Inf. Manag. (2021).
[14] V.K. Dixit, R.K. Malviya, V. Kumar, R. Shankar, An analysis of the strategies for overcoming digital supply chain implementation barriers, Decis. Anal. J. 10 (2024) 100389, http://dx.doi.org/10.1016/j.dajour.2023.100389.
[15] M. Momenitabar, Z.D. Ebrahimi, A. Abdollahi, W. Helmi, K. Bengtson, P. Ghasemi, An integrated machine learning and quantitative optimization method for designing sustainable bioethanol supply chain networks, Decis. Anal. J. 7 (2023) 100236, http://dx.doi.org/10.1016/j.dajour.2023.100236.
[16] R. Iqbal, F. Doctor, B. More, S. Mahmud, U. Yousuf, Big data analytics: Computational intelligence techniques and application areas, Technol. Forecast. Soc. Change 153 (2020) 119253.
[17] P. Chauhan, A. Mahajan, D. Lohare, Role of big data in retail customer-centric marketing, Natl. J. Multidiscip. Res. Dev. 2 (3) (2017) 484–488.
[18] F. Emmert-Streib, Z. Yang, H. Feng, S. Tripathi, M. Dehmer, An introductory review of deep learning for prediction models with big data, Front. Artif. Intell. 3 (2020) 4.
[19] N. Elgendy, A. Elragal, Big data analytics: A literature review paper, Adv. Data Min. Appl. Theor. Asp. 8557 (2014) 214–227.
[20] R.H. Hariri, E.M. Fredericks, K.M. Bowers, Uncertainty in big data analytics: survey, opportunities, and challenges, J. Big Data 6 (1) (2019) 1–16.
[21] K. Kambatla, G. Kollias, V. Kumar, A. Grama, Trends in big data analytics, J. Parallel Distrib. Comput. 74 (7) (2014) 2561–2573.
[22] J. Aversa, T. Hernandez, S. Doherty, Incorporating big data within retail organizations: A case study approach, J. Retail. Consum. Serv. 60 (2021) 102447.
[23] E. Aktas, Y. Meng, An exploration of big data practices in retail sector, Logistics 1 (2) (2017) 1–28.
[24] T.M. Choi, S.W. Wallace, Y. Wang, Big data analytics in operations management, Prod. Oper. Manage. 27 (10) (2018) 1868–1883.
[25] T. Van Nguyen, L. Zhou, A.Y.L. Chong, B. Li, X. Pu, Predicting customer demand for remanufactured products: A data-mining approach, European J. Oper. Res. 281 (3) (2020) 543–558.
[26] R. Basu, W.M. Lim, A. Kumar, S. Kumar, Marketing analytics: The bridge between customer psychology and marketing decision-making, Psychol. Mark. 40 (12) (2023) 2588–2611.
[27] A.S. Otto, D.M. Szymanski, R. Varadarajan, Customer satisfaction and firm performance: insights from over a quarter century of empirical research, J. Acad. Mark. Sci. 48 (3) (2020) 543–564.
[28] T. Stylianou, A. Milidis, The socioeconomic determinants of university dropouts: The case of Greece, J. Infrastruct. Policy Dev. 8 (6) (2024) 3729.
[29] H. Md Afnana, A. Shahriar, Y. Venkata, Revisiting customer analytics capability for data-driven retailing, J. Retail. Consum. Serv. 56 (2020) 1–13.
[30] B. Kitchens, D. Dobolyi, J. Li, A. Abbasi, Advanced customer analytics: Strategic value through integration of relationship-oriented big data, J. Manage. Inf. Syst. 35 (2) (2018) 540–574.
[31] Y. Zhao, X. Xu, M. Wang, Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews, Int. J. Hosp. Manag. 76 (2019) 111–121.
[32] J.A. Petersen, B.J. Paulich, F. Khodakarami, S. Spyropoulou, V. Kumar, Customer-based execution strategy in a global digital economy, Int. J. Res. Mark. 39 (2) (2022) 566–582.
[33] S.A. Otto, D.M. SzymanskiI, R. Varadarajan, Customer satisfaction and firm performance: insights from over a quarter century of empirical research, J. Acad. Mark. Sci. 48 (2020) 543–564.
[34] Kaggle, Supermarket dataset for predictive marketing 2023, 2023, [Online] Available at: https://www.kaggle.com/datasets/hunter0007/ecommerce-dataset-for-predictive-marketing-2023. (Accessed: 1 March 2023).
[35] A. Páez, G. Boisjoly, Exploratory data analysis, in: Discrete Choice Analysis with R, Springer International Publishing, Cham, pp. 25–64.
[36] M. Bata, Machine learning algorithms - A review, Int. J. Sci. Res. (IJSR) 9 (1) (2020) 381–386.
[37] M.M. Hassan, S. Zaman, S. Mollick, M.M. Hassan, M. Raihan, C. Kaushal, R. Bhardwaj, An efficient Apriori algorithm for frequent pattern in human intoxication data, Innov. Syst. Softw. Eng. 19 (1) (2023) 61–69.
[38] A. Kanavos, S.A. Iakovou, S. Sioutas, V. Tampakas, Large scale product recommendation of supermarket ware based on customer behaviour analysis, Big Data Cogn. Comput. 2 (2) (2018) 11.
[39] V. Grover, R.H. Chiang, T.P. Liang, D. Zhang, Creating strategic business value from big data analytics: A research framework, J. Manage. Inf. Syst. 35 (2) (2018) 388–423.
[40] P. Maroufkhani, R. Wagner, W.K. Wan Ismail, M.B. Baroto, M. Nourani, Big data analytics and firm performance: A systematic review, Information 10 (7) (2019) 226.
[41] D. Sjödin, V. Parida, M. Palmié, J. Wincent, How AI capabilities enable business model innovation: Scaling AI through co-evolutionary processes and feedback loops, J. Bus. Res. 134 (2021) 574–587.
[42] W. Wang, R. Arora, K. Livescu, J. Bilmes, On deep multi-view representation learning, in: International Conference on Machine Learning, PMLR, 2015, pp. 1083–1092.
[43] M. Aguiar, E. Hurst, Consumption versus expenditure, J. Political Econ. 113 (5) (2005) 919–948.
[44] X. Jaravel, The unequal gains from product innovations: Evidence from the us retail sector, Q. J. Econ. 134 (2) (2019) 715–783.
[45] S. Athey, G.W. Imbens, Machine learning methods that economists should know about, Annu. Rev. Econ. 11 (1) (2019) 685–725.
[46] G. James, D. Witten, T. Hastie, R. Tibshirani, J. Taylor, Statistical learning, in: An Introduction to Statistical Learning: With Applications in Python, Springer International Publishing, Cham, 2023, pp. 15–67.
[47] X. Shu, Y. Ye, Knowledge discovery: Methods from data mining and machine learning, Soc. Sci. Res. 110 (2023) 102817.
[48] I.A. Ajah, H.F. Nweke, Big data and business analytics: Trends, platforms, success factors and applications, Big Data Cogn. Comput. 3 (2) (2019) 32.
[49] S. Tasos, M.I. Amjad, M.S. Awan, M. Waqas, Poverty alleviation and microfinance for the economy of Pakistan: A case study of Khushhali Bank in Sargodha, Economies 8 (3) (2020) 63.