Available at e-Journal Universitas Sultan Ageng Tirtayasa

JOURNAL INDUSTRIAL SERVICESS

journal homepage: http://jurnal.untirta.ac.id/index.php/jiss

eISSN 2461-0631

Original research

Designing an analytics dashboard for knowledge extraction in the retail industry using descriptive and predictive analytics

Muhammad Aldyan Ruansyah, M Zaky Hadi*, Muhammad Iqbal, Juniwati Juniwati, Lina Aulia

Department of Industrial Engineering, Institut Teknologi Sumatera, Indonesia

ARTICLE INFO

Article history:
Received 29 July 2024
Received in revised form 5 December 2024
Accepted 6 December 2024
Published online 30 December 2024

Keywords:
Big data analytics
Descriptive analytics
Knowledge extraction
Predictive analytics
Retail industry

Editor:
Bobby Kurniawan

Publisher's note:
The publisher remains neutral concerning jurisdictional claims in published maps and institutional affiliations.

ABSTRACT

With the rise of digitalization and data growth, many business sectors are increasingly focused on data. As a result, developing a descriptive and predictive analytics dashboard is essential for extracting insights from complex datasets and supporting decision-making. This study addresses data analytics challenges in the retail industry by designing a dashboard system for analyzing and visualizing data. It also proposes future predictions for customer analytics, simplifying the interpretation of results from complex datasets obtained through business activities in the retail supply chain. These processes have implications for decision-making in areas such as customer segmentation, demographic analysis, and sales performance. The dashboard design includes a pre-processing stage for retail industry datasets and the creation of descriptive and predictive analytics models using clustering methods. It also simulates the development of these models into a comprehensive dashboard. The primary analysis employs the K-Means algorithm for RFM (Recency, Frequency, Monetary) analytics, customer segmentation, and demographic analysis. Results show that the use of this dashboard enhances the visualization of data, supporting decision-making processes related to marketing strategies, sales performance, and inventory management. By applying advanced analytic techniques to marketing strategies, supply chain management, and inventory planning, retail businesses can optimize their operations. This optimization is achieved through better data analytics tools.

1. Introduction

The global economy relies heavily on the retail industry, which is dynamic and essential worldwide. By 2023, global retail sales were expected to reach around $27 trillion, highlighting its significant impact on the economy. Over the years, this sector has experienced substantial growth due to factors such as rising consumer spending, advancements in technology, and the emergence of online business. Online sales alone account for almost 20% of total retail sales. Retailing is a fast-moving industry with a constantly changing face, owing to shifts in consumer preferences, the inclusion of digital systems, and the growth of omnichannel networks. Consequently, firms within this sector are increasingly turning to data analytics to understand market trends, optimize operations, and improve the customer experience. This helps them stay relevant amid tough competition. [1], [2]. However, the growth and diversity of the retail industry are accompanied by pervasive challenges. These challenges primarily involve managing complex datasets that are not straightforward to interpret [3].

Retailers collect a high volume of data from thousands of sources, including sales transactions, customer feedback, inventory records, and online interactions. Often, this data is considered "dirty" due to its heterogeneity, large volume, and high velocity of generation. This makes processing and analysis highly ineffective [4], [5], [6]. Moreover, the use of multi-channel information—such as data from on-site and online platforms—further complicates the situation.

Data should be securely managed by retailers to maximize the accuracy of value extraction. Advanced analytical techniques, including machine learning, are essential for turning complex data into actionable strategies through predictive modeling. A thorough understanding of customer behavior can lead to improvements in supply chain decisions, helping retail businesses grow by increasing consumer satisfaction [4], [5], [6], [7], [8].

*Corresponding author:
Email: zaky.hadi@ti.itera.ac.id

http://dx.doi.org/10.62870/jiss.v10i2.27900

It is essential to use big data analytics to overcome the dataset challenges faced by retailers. Big data analytics is a necessity today for addressing the challenges posed by datasets for any retailer. Retailers who adopt advanced analytics can transform datasets of any scale into insights, uncovering patterns and correlations that drive efficient decision-making [9], [10], [11]. This capability empowers organizations to make data-driven decisions and improve operations. Personalized customer relationships can also be facilitated, along with marketing enhancements at the click of a mouse. For example, predictive analytics can anticipate demand, ensuring the organization holds the correct amount of stock, thus preventing stockouts and overstocking. Clustering algorithms allow retailers to target customer groups more effectively, improving campaign success and fostering loyalty. Sentiment analysis provides valuable product feedback, guiding developers to areas that require improvement in design. Big data analytics simplify complex datasets, enabling retailers to stay competitive by quickly adapting to market changes and consumer preferences [12], [13], [14]. Ultimately, big data analytics are crucial for transforming large retail datasets into strategic resources that drive revenue and innovation [15], [16], [17].

Big data analytics in retail currently focuses on techniques for leveraging massive datasets in business operations. Studies have found that machine learning and deep learning models can successfully predict customer behavior, enabling the adoption of personalized recommendations, which leads to increased sales and improved customer engagement [18], [19]. Procurement and supply chain operations are changing with the use of real-time process data analytics. Primarily, predictive analytics focuses on identifying areas where cost savings can be made in delivering supply chain services. In the next five years, this field is expected to grow exponentially, as it becomes increasingly effective on the scale [20], [21]. Another important area is sentiment analysis of customer reviews and social media interactions, which informs marketing strategies driven by customer sentiments [22]. More specifically, prior research has studied the integration of big data with Internet of Things (IoT) technology to enhance in-store customer experiences and inventory management [23], [24]. Together, these advances highlight the potential of big data analytics to completely reshape retail strategy and execution. While research on predictive algorithms is ongoing, integrating descriptive and predictive analytics to build a retail decision-making dashboard remains underexplored. Therefore, developing a dashboard that combines descriptive and predictive analytics to support real-time decision-making is highly valuable.

This study aims to establish a model called the Retail Analytics Dashboard, which will handle complex datasets from the retail industry. The effort seeks to apply descriptive and predictive analytics to create a roadmap for decision-making through knowledge extraction. It involves the preprocessing design of retail datasets and the development of analytical models for deployment in dashboards. With such structured research, all aspects are thoroughly analyzed, ensuring practical applications.

Unlike recent studies that rely on a single predictive method, this study adopts an integrated model that combines RFM clustering with predictive analytics. These insights can be displayed in a user-friendly dashboard to facilitate actionable plans for marketing, sales, and inventory management. Therefore, this work contributes to the sophistication of targeting strategies and the reduction of complexities in applying big data analytics to retail. Marketing optimization, supply chain management, and inventory planning could all benefit from increased competitiveness and operational efficiency.

2. Material and method

2.1. Research flow

Several stages are included in the data analytics research flow, as shown in Fig. 1 [25], [26], [27]. The process begins with understanding the data and analyzing the structure of the retail dataset.

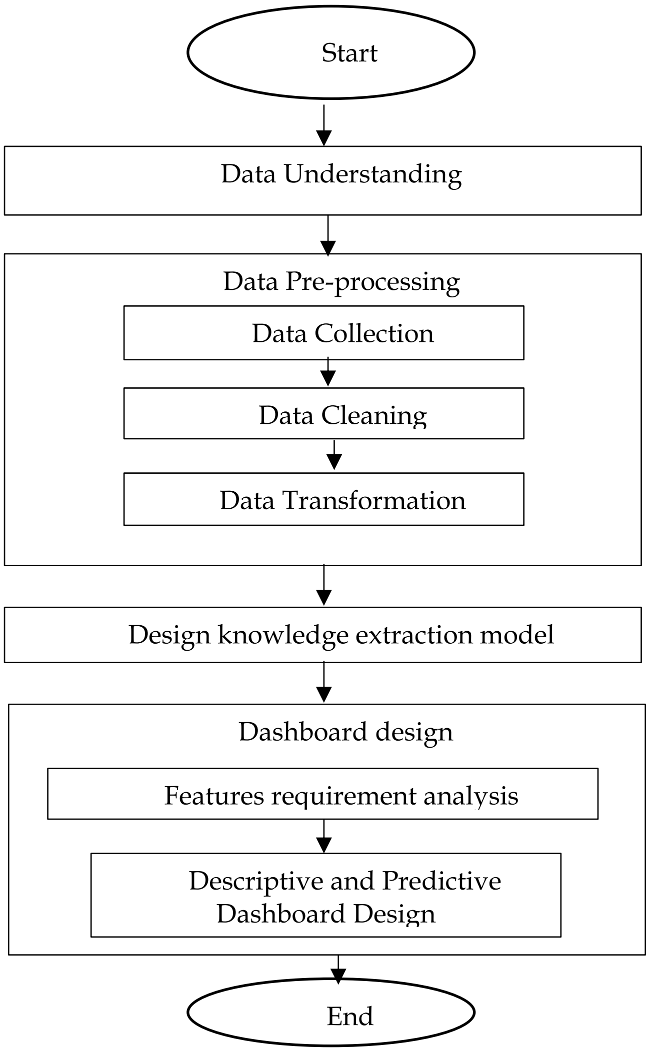

Figure 1. Research flow

It then moves on to preprocessing, which involves gathering, cleaning, and transforming the datasets into high-quality, relevant data for analysis. The next stage involves both descriptive and predictive analytics, facilitated by clustering algorithms. Following this, a model will be designed using descriptive and predictive analytics, culminating in a user-friendly dashboard to support improved decision-making. This structured approach ensures that each stage builds on the previous one, resulting in a comprehensive model that enhances strategic decision-making in retail.

2.2. Data understanding

Big data analytics itself has a phase for data understanding, which is quite an important aspect. It involves eliciting the data completely to ensure that it is completely fit for analysis. It also includes Exploratory Data Analysis (EDA) to extract patterns, trends, outliers, or anomalies by statistical summaries and visualization. Including domain knowledge in this area permits interpretation and validation in context. Such a profound understanding ensures the making of well-initial informed choices and proper data preparation. Therefore, allowing the reliable and accurate development of models increases the relevance of the analytics [28], [29].

This paper uses the "Consumer Behavior and Shopping Habits Dataset" from Kaggle to develop an analytics dashboard model for complex retail industry datasets [30]. The "Consumer Behavior and Shopping Habits Dataset" on Kaggle includes various facets of a customer's life, such as demographics like age, sex, and location. It comprises transactional data, such as purchasing behaviors or clusters of goods purchased over time as a function of spending habits. It allows retailers to identify the items that have had the best sales performance through time among specific demographic segments or groups. This information also indicates changes in the patterns among some of these groups that could affect target products' profitability for a business.

A retail company might be interested in evaluating its advertising techniques by investigating the relationship between purchasing preferences and age, gender and geographical features among others of consumers. Through this method, it is possible to identify how such factors influence people’s choices while making purchases. This dataset is crucial for retailers, researchers, and analysts studying consumer behavior to enhance marketing strategies and improve customer satisfaction in the retail industry.

2.3. Data pre-processing

Retail data analytics require data pre-processing, which is a key step to convert raw data into a clean and useful format. This process includes various processes such as data cleaning (handling missing values), transformation, integration and dimensionality reduction. It ensures that the data is ready for analysis including meaningful insights and informed decisions.

The maintenance of data quality and dependability is crucial in data preprocessing through Data Cleaning (handling missing values, inconsistency, duplicates, and outliers). In this research, missing values were addressed through the elimination of rows or columns that exhibited anomalies or using statistical methods such as mean, median, or mode depending on the type of data attributes. Furthermore, it addressed the possible errors like different formats and units standardized or unified data entries into one format. A further motive in deleting duplicates is to guarantee the uniqueness of every data point to avoid biased analysis results. It is finally possible to identify outliers that could skew results from statistical analysis using methods such as Z-score or IQR. Sometimes, it is simply removing them or transforming them in accordance with their context and the nature of the data. These involve increasing accuracy, consistency, and every quality feature in a dataset for retail data analytics preparation [29], [31].

Normalization, categorical encoding and feature construction (this is essentially a form of transformation) are critical processes in data pre-processing/preparation. These techniques are employed to enhance the performance of machine learning models. The process of integrating disparate data sources to ensure accuracy is referred to as integration. Conversely, reduction (which serves a distinct purpose) minimizes a data set by selecting a limited number of pertinent characteristics. However, this reduction does not imply a denial of significant information; rather, it aims to streamline the data while preserving its essential value [32]. This research adopted One-Hot Encoding transformation to make data available for analysis and clustering models to process them in the form of data vectors. The process of One-Hot Encoding modifies categorical features into binary vector columns for further analysis. The conversion of categorical data into binary vector maps enhances the accuracy of the model, interpretation, and calculations.

2.4. Knowledge extraction model

Predictive analytics model searches for actionable insights and patterns from which future predictions are derived. After the data has been collected and preprocessed , there is selection of important factors that, along with data mining processes, will enable the surfacing of hidden patterns. The knowledge extracted would be used to build and train its predictive models, evaluated apparently by the accuracy or through such metrics as Root Mean Squared Error (RMSE). Interpretation of the knowledge thus extracted would provide further understanding of how different features and patterns contribute [31], [33].

RFM clustering is a powerful model in retail analytics. When customers are analyzed based on RFM—Recency (how recent the last purchase was), Frequency (how often purchases occur), and Monetary (how much a customer spends over time)—they can be clustered using K-means on RFM-based metrics into four distinct groups: 'loyal', 'big spenders', 'inactive', and possibly others. This segmentation allows marketing efforts to become more targeted and experiential, creating personalized customer experiences and improving reference management, which leads to greater sales and better customer retention. It is through such RFM clustering that retailers can effectively target the right customers with their marketing campaigns, enhancing customer satisfaction [31], [33].

The model adopted for knowledge extraction in this research is a clustering-based knowledge extraction model, which exploits similarity in behavior within a complex retail dataset. It is a predictive analytics method that groups similar data points before applying models to forecast future outcomes. Clustering identifies intrinsic groupings through its unsupervised learning technique, based on characteristic features. Once formed, separate predictive models, such as regression, classification, or even time-series prediction, are applied to analyze each group exclusively, leveraging the unique characteristics of each cluster. This technique improves the accuracy of predictions by considering the distinct patterns within each cluster, leading to more fine-grained and actionable insights.

The K-Means clustering algorithm is a popular method for partitioning a dataset into k distinct, non-overlapping subsets (clusters). The goal is to minimize the variance within each cluster. The objective of *K-Means clustering* is partition n data points $x_1, x_2, \dots, x_n$ into k cluster $C_1, C_2, \dots, C_k$. Each cluster has a centroid μ_i, which is the meaning of all points in the cluster is defined in Eq. (1), where $|C_i|$ is the number of points in cluster C_i. The objective function is to minimize the sum of squared distances between data points and their respective cluster centroids as shown in Eq. (2), where $||x - \mu_i||^2$ is the squared Euclidean distance between data point x and centroid μ_i.

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{1}$$

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

$$C_i = \{x_j : ||x_j - \mu_i||^2 \leq ||x_j - \mu_l||^2 \ \forall l \in \{1,2,\dots,k\} \tag{3}$$

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{4}$$

The K-means algorithm is conducted using the procedure as follows.

1. Randomly select k initial centroids from the data points.
2. Assign each data point to the nearest centroid based on the Euclidean distance using Eq. (3).
3. Recompute the centroid of each cluster using Eq. (4).
4. Repeat steps 2 and 3 until the centroids no longer change or change below a specified threshold.

2.5. Dashboard design

The design and development of dashboards that require retail analytics at a large-scale focus on creating a visual interface that makes operational changes easier. Some of the most critical metrics include RFM, sales performance, customer behavior, stock levels, and marketing effectiveness. By utilizing data from point-of-sale systems, available customer relationship management and supply chain management, accuracy and standardization in retail analytics are achieved. The aim of the design is to enable users of the dashboard to have pleasant experience and easy interaction with the data [30].

K-Means computation is performed using Python and Power BI, which are then used to create dashboards that transform raw data into actionable insights. Data modeling involves importing, transforming, and establishing relationships among several data tables. DAX (Data Analysis Expressions) will be used to calculate key metrics and other KPIs. Various visual elements, such as bar charts, line charts, pie charts, heat maps, and gauges, are included in the dashboard to display metrics on sales, consumer insights, inventory, and marketing analysis. For example, users can customize the view using slicers, filters, and drill-down options for more detailed analysis. These features ensure the smooth performance of the dashboard and allow for its deployment to Power BI Service for sharing, collaboration, and scheduled data refreshes. This fact-based approach supports decision-making in businesses, providing valuable insights across all retail operations.

3. Results and discussions

3.1. Datasets

Kaggle hosted an open public comprehensive dataset on consumer behavior and shopping habits in [30], described in detail from the Data Understanding section. Demographic information, purchase history, product preferences and shopping frequency are just some of the variables that make this dataset complete. The dataset on "Consumer Behavior and Shopping Habits" is a comprehensive dataset of retail consumer data, which is an important resource in market analysis and a basis for personalized marketing. The datasets include Customer IDs for identifying the customers, Age and Gender for covering the demography, and Purchase Amount in USD, which has transaction values. The Dataset holds information on Item Purchased, Category, and Location that provides the insights of product preferences per state. Preference-based dimensions such as Size, Color, and Season save from bringing down to specific consumer choices.

Table 1.
Data understanding and cleaning process

No	Attributes Name	Total Record/ row	Data Type	Missing values identified
0	Customer ID	3900	Int64	0
1	Age	3900	Int64	0
2	Gender	3900	Object	0
3	Item Purchased	3900	Object	0
4	Category	3900	Object	0
5	Purchase Amount	3900	Int64	0
6	Location	3900	Object	0
7	Size	3900	Object	0
8	Color	3900	Object	0
9	Season	3900	Object	0
10	Review Rating	3900	Float64	0
11	Subscription Status	3900	Object	0
12	Payment Method	3900	Object	0
13	Shipping Type	3900	Object	0
14	Discount Applied	3900	Object	0
15	Promo Code Used	3900	Object	0
16	Previous Purchases	3900	Int64	0
17	Preferred Payment Method	3900	Object	0
18	Frequency of Purchases	3900	Object	0

Note: 3900 entires with 19 columns / attributes (3900 x 19 matrix data)

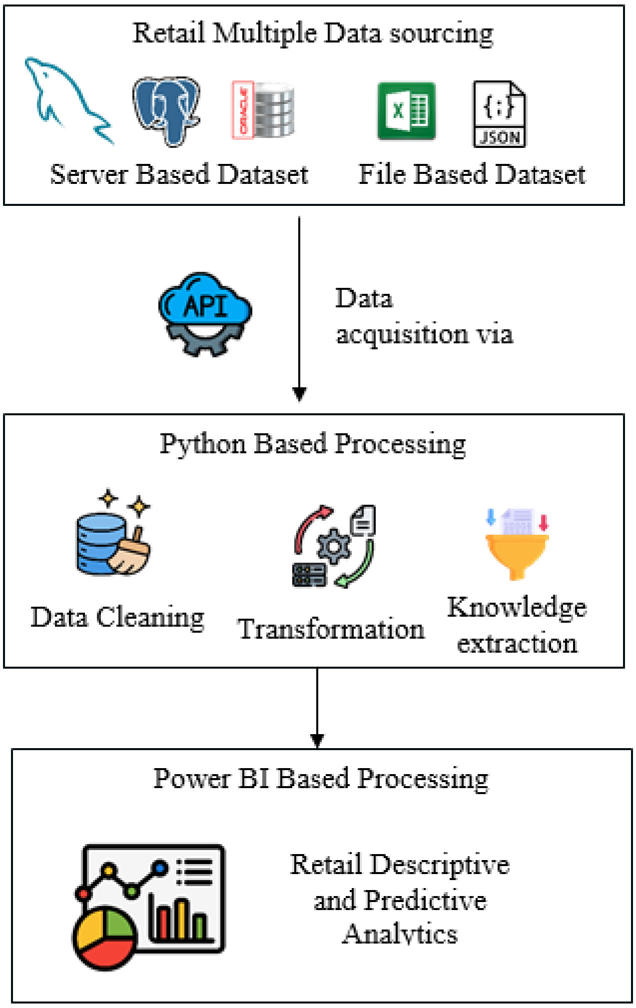

Figure 2. The retail analytics framework

These features are important for deep analysis at both levels, making improvements to the shopping experience of consumers in time to come while refining retailers' marketing strategies. The 19 columns by 3900 rows high-dimensional data set contain both numerical and categorical data types: thus, it is complicatedly complex to be analyzed using regular methods of data analysis. Therefore, advanced data analysis techniques are needed to extract any meaningful insights at all from it. Below is a structural output sheet which shows "Consumer Behavior and Shopping Habits Dataset" identified using python with Pandas and NumPy Python Library as shown in Table 1.

Table 2.
Data transformation process

Index	Customer ID	Age	…..	Frequency of Purchases
0	1	55	…..	0
1	2	19	…..	0
2	3	50	…..	0
3	4	21	…..	0
4	5	45	…..	1
5	6	46	…..	0
6	7	63	…..	0
7	8	27	…..	0
8	9	26	…..	1
9	10	57	…..	0

Note: Found: Attributes 'Item Purchased' contains non-numeric values that cannot be converted to integers. Shown 10 of 3900 datasets for example

Table 3.
Clustered season

Index	Customer ID	Season	Item Purchased	Category	Cluster
0	1	3	2	1	0
1	2	3	23	1	1
2	3	1	11	1	2
3	4	1	14	2	2
4	5	1	2	1	0
5	6	2	20	2	1
6	7	0	16	1	2
7	8	3	18	1	1
8	9	2	4	3	0
9	10	1	7	0	0

Table 4.
Clustered location

Index	Customer ID	Season	Item Purchased	Cluster
0	1	16	2	2
1	2	18	23	1
2	3	20	11	1
3	4	38	14	0
4	5	36	2	0
5	6	49	20	0
6	7	25	16	1
7	8	17	18	1
8	9	47	4	0
9	10	24	7	1

Table 5.
Clustered shipping type

Index	Customer ID	Season	Shipping type	Location	Cluster
0	1	3	1	16	1
1	2	3	1	18	1
2	3	1	2	20	1
3	4	1	3	38	0
4	5	1	2	36	0
5	6	2	4	49	0
6	7	0	2	25	1
7	8	3	2	17	1
8	9	2	1	47	0
9	10	1	0	24	1

In the construction of the dashboard model, we design a systematic process that employs the use of both analytical and visualization tools for smoothening the development between the data analytics and the end-user interface as shown in Fig. 2. The main software involves Python for pre-processing and clustering data then designing the dashboard using Power BI. An example of the preprocessing steps that took place was cleaning, transforming and clustering the data using python libraries, such as Pandas, NumPy, and Scikit-learn, to prepare the data that was uploaded from the structured dataset present on Kaggle into raw data.

Table 6.
Clustered gender

Index	Customer ID	Gender	Item Purchased	Color	Cluster
0	1	1	2	7	1
1	2	1	23	12	2
2	3	1	11	12	1
3	4	1	14	12	2
4	5	1	2	21	0
5	6	1	20	23	0
6	7	1	16	7	2
7	8	1	18	4	2
8	9	1	4	19	0
9	10	1	7	16	0

In Fig. 2, a dataset is being retrieved and inserted into a local storage for the purpose of analysis; alternatively, a server or API could be used for building a data pipeline to facilitate real-time updates. The structure of the database can be defined through the help of any relational database management system or document-based system file such as MySQL or SQLite or an Excel file. This setup allows defined structured querying which in turn ensures possible and simple data manipulation and analysis.

The front-end dashboard was constructed with Power BI for rich data visualizations. Core metrics that concern customer segments and sales performance were processed into Power BI through Data Analysis Expressions (DAX) calculated fields and metrics. Intuitive insights were developed by incorporating popular visual graphical analysis instruments bar, line, and heatmap which ensured display of data meaningfully. The interactivity features such as slicers and drill down enabled flexible exploration of data that would suit the interest of various stakeholders. Such an informational integration of the two features guarantees scalability and well-suited solutions for retail analytics.

Considering the significance of numeric values—such as those derived from customer ID, age, review ratings, prior purchases and purchase frequencies — other vital numeric factors (in this context) revolve around floating-point values, particularly product prices. Important categorical variables are gender, types of items purchased, product categories, areas, color, size, season, subscription status, payment mode, shipping mode, and percentages on discounts. The dataset would thus avail diversity to interesting descriptive and predictive analytics that could help present trends and probably predict behavior to aid strategic decision-making in retail.

This study examines products indirectly during the data understanding phase, by focusing transactional and demographic variables, for example, frequency of purchase, monetary value, and preferences of customers rather than product characteristics like perishability or obsolescence. Products per se are not that critical as the study's objective in this paper is to make enhanced decision making through customer segmentation analytics rather than inventory analytics. This dashboard then produces actionable insights in terms of customer behavior and sales trends through the exploitation of clustering and RFM metrics that are generally meaningful regardless of what specific product attributes are. This ensures flexibility and applicability of the model in retail customer analytics, making the model useful in all retail environments without restriction to the specifics of what the product nature is like.

Data cleaning is followed by data transformation aimed at making the dataset ready for analysis. In this process, the columns may be merged or split, and new features added. The main purpose of data transformation is to reduce the complexity of the dataset and ensure that it fits into the analytical aim of the research. Moreover, the data can be reformatted into forms that are easier to manipulate and analyze using Visual Studio Code or Python. This phase ensures that the data is not only clean but also best structured for subsequent stages of descriptive as well as predictive analytics.

The transformation of data as illustrated in Table 2 involves some stages that require additional scrutiny and upgrading of the data set (first 10 data for illustration). In the first place, changing the data into text ensures that there are no errors. Such a process enables proper recognition and treatment of all non-numeric components of the dataset. The next step is to recode textual information back into numbers to confirm successful conversion. As such, an error handling device is used to deal with undetectable numerical figures that cannot be turned into different versions thus maintaining these records on its original content.

In the next phase, we use one-hot encoding to convert categorical variables into numbers. It transforms categorical data into binary vectors, each unique category being represented by a separate column containing binary values. The data becomes more appropriate for mathematical calculations, thereby increasing modeling accuracy. This research outputs the successful conversion of categorical variables into a numerical format suitable for predictive analytics.

3.2. Knowledge extraction model

The K-Means clustering analysis in this research goes through several critical phases. A dataset is loaded using Pandas for data preparation, and LabelEncoder is employed to convert categorical variables into numerical ones, as previously described. Selected features for clustering are then placed in a separate dataframe. The K-Means model is created by specifying a set number of clusters—in this case, three—and setting a random seed to make the process reproducible. The data clustering process categorizes the data based on variables such as "Season," "Category," and "Item Purchased," among others.

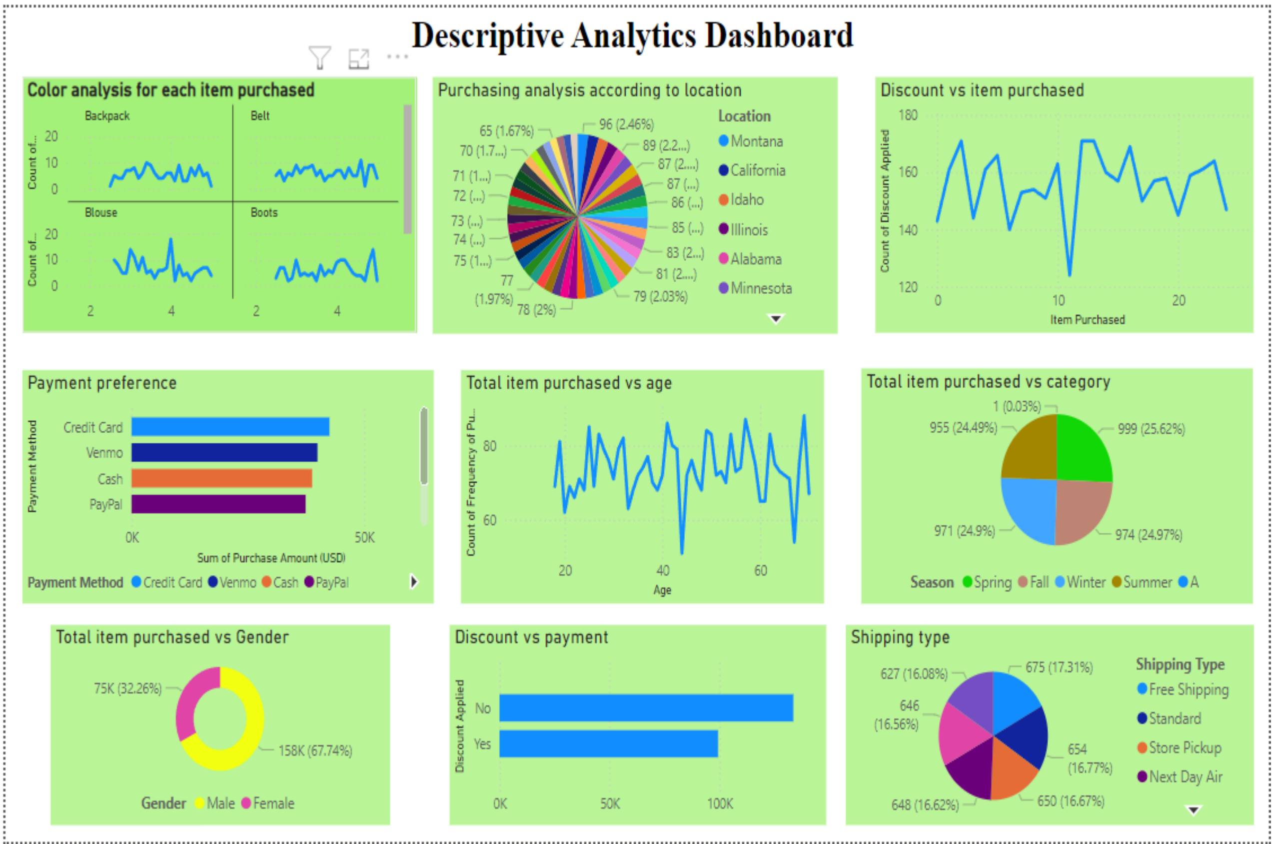

Figure 3. Descriptive dashboard analytics

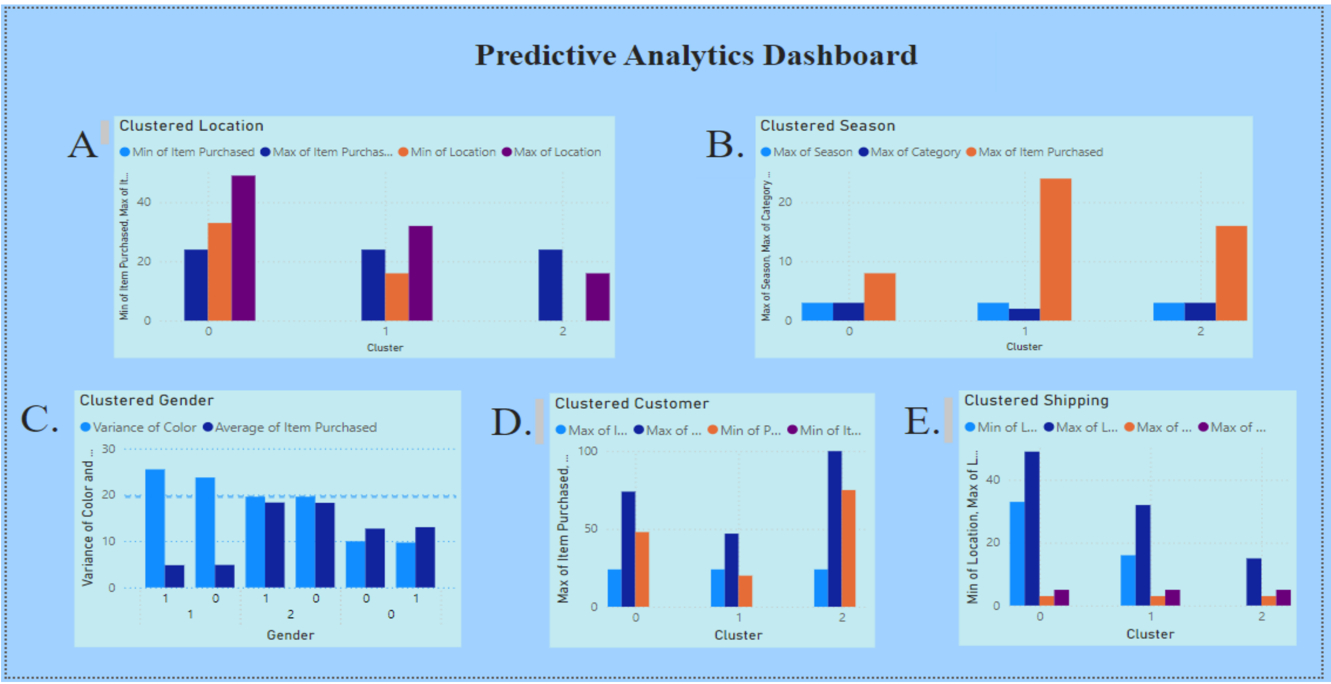

Figure 4. Predictive Dashboard Analytics

The K-Means clustering has produced data results in various new data frames, each corresponding to different clustering criteria, such as “Season” (Table 3), “Location” (Table 4), “Shipping Type” (Table 5), and “Gender” (Table 6). These data frames contain variables like those in the original dataset and are saved as Excel files with names that correspond to their clustering criteria, such as “Clustered_Season” and “Clustered_Location.” Additionally, this study also includes the specific code for each clustering scenario, from data preparation to the final generation of clustered data. The groupings from this exploration yielded unambiguous findings for each category, which are: "Season," "Location," "Shipping Type," and "Gender." In the case of the "Season" clustering, we classified the data into three clusters: "Season," "Category," and "Item Purchased." The results indicated that cluster 0 represented the winter season, where there was a high level of variability in item purchases (ranging from 0 to 23) and the number of categories (ranging from 0 to 3). Cluster 1, representing the summer season, showed less variation in these two variables, with item purchases ranging from 17 to 24 and categories ranging from 0 to 2. The spring cluster represented items purchased (ranging from 9 to 16) across all categories (0 to 3). Through this method, three distinct clusters were identified. Regarding the "Location" cluster, it was based on the analysis of two variables: "Location" and "Item Purchased." Thus, the inclusion of "Season," "Location," and "Shipping Type" in this clustering process resulted in the creation of a new variable labeled "Cluster." This analysis revealed specific shipping preferences depending on the season and location. Finally, "Gender" constituted another cluster, which incorporated three additional variables: "Gender," "Color," and "Item Purchased." These findings were exported to Excel files, with each individual file containing the basic variables mentioned above, along with the clustering information.

The clustering analysis, which segments the dataset into various groups based on specified attributes, reveals trends and relationships. The K-Means machine learning algorithm has further converted categorical variables into numerical values to classify data for predictive analytics and decision-making. These clustered DataFrames could serve as the foundation for further analyses and graphical representations, thus contributing to the larger goals of this research.

The analytic dashboard is divided into three sections: top, middle, and bottom, as shown in Fig. 3. The descriptive and predictive analytics involve the extraction of knowledge from the sample dataset, “Consumer Behavior and Shopping Habits Dataset.” According to the analysis presented in the top section, factors such as rating, location, and discount affect product purchases. For instance, among all reviewed products, backpacks with a rating of 3.40 had the highest sales, but only one product scored above four out of five. Additionally, this section provides insights into how different products perform under various conditions.

According to Fig. 4 for predictive analytics, the dashboard employs initial modeling in Python for generating predictive clusters. The results are then used as inputs for further modeling in Power BI, where it is displayed in an easy-to-use manner. This allows stakeholders to make well-informed decisions based on both past events and future developments since it combines descriptive and predictive analytics. This combination of methods helps retailers predict customer preferences, track market trends, and adjust inventories, improving competitiveness and customer satisfaction.

These insights come to the next level with the use of predictive analytical techniques to predict future customer behavior. The dashboard looks into past data and predicts the customers who are going to buy repeatedly, the ones likely to churn, and segments that are more receptive to certain types of marketing. This forecasting ability enables companies to predict issues, effectively allocate marketing budgets and put measures in place to retain customers from being lost. RFM clustering added into the dashboard becomes a significant tool in strategizing and managing customer relationships in that respect.

The constructed dashboard will indeed have major managerial implications regarding decision-making, costing it into the retail industry, as it becomes a comprehensive data-driven tool to interpret customer behavior and sales trends. Managers can deploy descriptive and predictive analytics techniques to segment customers through RFM analysis, project demand patterns for high-value customer groups, and develop a marketing strategy targeting these patterns, which will help retain customers. By this method, the dashboard employs interactive visualizations to provide clarity in interpreting complicated data rather than manual analyses and accelerates decision making. The possibility of real-time data makes it possible for managers to act quickly since the market situation has changed. Optimize inventory levels and minimize wastage. The delivery thus increases operational efficiency, better resource allocation, and ultimately, very competitiveness in this rapidly transforming retail landscape.

4. Conclusions

The research findings are summarized, and its major contributions are outlined in this conclusion. This research successfully completed data pre-processing, including cleaning and transformation, to prepare the dataset for retail analytics modeling. Clustering algorithms were applied to the pre-processed data, creating descriptive and predictive models by segmenting it into three clusters based on various variables. An optimized dashboard was created to visualize descriptive trends and predictive insights, aiding in sales pattern recognition and customer segmentation for retail analytics.

However, even though this research provides a structured method of analyzing and visualizing complex retail business dataset, there are certain limitations that were acknowledged including product characteristics analytics for the dashboard. Therefore, it is recommended that future research should include product characteristics like obsolescence and perishability into the inventory analytics framework with a view to improving decision making in several retail operations. For example, by integrating shelf-life, product lifecycle stages and other time-sensitive factors into the clustering models and predictive dashboards, it would make inventory management much more efficient and less wasteful. Stock rotation algorithms for perishables or obsolescence prediction models for high-tech stocks would make it possible for retailers to customize strategies for different product categories. It may also create room for real-time alerting or decision support tools for the management of time-sensitive inventory making analytics respond to the operational needs more closely.

Encouragement of the development of prescriptive analytic research involves the application of computational intelligence in warehousing and logistics to take better management decisions. Its appropriate integration into machine learning and optimization algorithms will greatly improve inventory management, demand forecasting, and route planning as shown in. Real-time data and the Internet of Things (IoT) devices could also improve dynamic decision-making and, hence, reduce costs and increase efficiency in the entire supply chain management process. Future research may deal with advanced simulation techniques or models that would predict and prescribe optimal actions in very complex real-world logistics systems, such as demand variability, transportation delays, or warehouse capacity limitations.

Declaration statement

Muhammad Aldyan R. and M Zaky Hadi (Same Contribution): **Conceptualization, Methodology, Writing-Original Draft, Calculation and Data Collection, Writing-Review**. Muhammad Iqbal, Juniwati, and Lina Aulia: **Conceptualization**, **Editing.**

Acknowledgement

We sincerely thank everyone who contributed to the success of this research in big data analytics. We also extend our appreciation to Optimization and System Design Laboratory (POSI) Laboratory and Faculty of Industrial Technology, Institut Teknologi Sumatera for supporting the administrative aspects of this research.

Disclosure statement

The author asserts that this manuscript has no conflict of interests, and it has been processed according to relevant journal's provisions and policies, to avoid deviations from the codes of publication ethics in other forms.

Funding statement

The authors received no funding for this research.

Data availability statement

The authors confirm that the data supporting the findings of this study are available within the article or its supplementary materials.

AI Usage Statement

Generative AI (and AI-assisted tools) were utilized to enhance the clarity and readability of certain sections of this manuscript, especially those deemed ambiguous or potentially challenging for readers to comprehend. The authors have meticulously reviewed and refined all AI-generated material to guarantee its accuracy and consistency with the research objectives. Full responsibility for the scientific content, conclusions and integrity of the manuscript remains with the authors; and the authors disclose the use of AI to ensure transparency and adherence to publisher guidelines.

References

[1] Shopify Plus, "Commerce Trends 2023: Industry Report," 2023.

[2] Global Base Sourcing Limited, "2023 Statistics, Trends, and Forecasts for the Retail Industry," https://globalbasesourcing.com/2023-retail-industry-statistics-trends-forecasts.

[3] P. C. Verhoef, P. K. Kannan, and J. J. Inman, "From Multi-Channel Retailing to Omni-Channel Retailing. Introduction to the Special Issue on Multi-Channel Retailing.," *Journal of Retailing*, vol. 91, no. 2, 2015, doi: 10.1016/j.jretai.2015.02.005.

[4] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS Q*, vol. 36, no. 4, 2012, doi: 10.2307/41703503.

[5] S. F. Wamba, A. Gunasekaran, S. Akter, S. J. fan Ren, R. Dubey, and S. J. Childe, "Big data analytics and firm performance: Effects of dynamic capabilities," *J Bus Res*, vol. 70, 2017, doi: 10.1016/j.jbusres.2016.08.009.

[6] A. Frederiksen, "Competing on analytics: The new science of winning," *Total Quality Management & Business Excellence*, vol. 20, no. 5, 2009, doi: 10.1080/14783360902925454.

[7] R. Bose, "Advanced analytics: opportunities and challenges," *Industrial Management & Data Systems*, vol. 109, no. 2, 2009, doi: 10.1108/02635570910930073.

[8] O. Kwon, N. Lee, and B. Shin, "Data quality management, data usage experience and acquisition intention of big data analytics," *Int J Inf Manage*, vol. 34, no. 3, 2014, doi: 10.1016/j.ijinfomgt.2014.02.002.

[9] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods," *J Bus Res*, vol. 70, 2017, doi: 10.1016/j.jbusres.2016.08.001.

[10] H. V. Jagadish, "Big Data and Science: Myths and Reality," 2015. doi: 10.1016/j.bdr.2015.01.005.

[11] S. Fosso Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study," *Int J Prod Econ*, vol. 165, 2015, doi: 10.1016/j.ijpe.2014.12.031.

[12] M. T. Sembiring, K. Adhinata, D. Wahyuni, and M. Z. Hadi, "Determining Kansei Words in Chocolate Product Development Model Design Based on Social Media Trend by Using Key Element Extraction (KEE) Algorithm," in *IOP Conference Series: Materials Science and Engineering*, 2019. doi: 10.1088/1757-899X/505/1/012031.

[13] M. Z. Hadi, T. Djatna, and Sugiarto, "A modeling of dynamic storage assignment for order picking in beverage warehousing with Drive-in Rack system," in *IOP Conference Series: Materials Science and Engineering*, 2018. doi: 10.1088/1757-899X/337/1/012020.

[14] M. Z. Hadi, "Peluang Implementasi Teknologi Big Data Dan Block Chain Untuk Peningkatan Kinerja Perdagangan Pada Sektor UMKM di Indonesia Pada Era Industri 4.0," *Cendekia Niaga*, vol. 3, no. 1, 2020, doi: 10.52391/jcn.v3i1.463.

[15] F. Chaffey, D., & Ellis-Chadwick, "Digital Marketing: Strategy, Implementation and practice (6th ed.)," *International Journal of Business and Economic Affairs*, vol. 6, no. 2, 2016.

[16] D. R. Heath, "Prediction machines: the simple economics of artificial intelligence," *Journal of Information Technology Case and Application Research*, vol. 21, no. 3–4, 2019, doi: 10.1080/15228053.2019.1673511.

[17] J. Manyika, M. Chui Brown, B. B. J., R. Dobbs, C. Roxburgh, and A. Hung Byers, "Big data: The next frontier for innovation, competition and productivity," *McKinsey Global Institute*, no. June, 2011.

[18] Z. Chen, J. Zhao, and C. Jin, "Business intelligence for Industry 4.0: predictive models for retail and distribution," *International Journal of Retail and Distribution Management*, 2023, doi: 10.1108/IJRDM-02-2023-0101.

[19] J. Tian, Y. Zhang, and C. Zhang, "Predicting consumer variety-seeking through weather data analytics," *Electron Commer Res Appl*, vol. 28, 2018, doi: 10.1016/j.elerap.2018.02.001.

[20] H. N. Dai, H. Wang, G. Xu, J. Wan, and M. Imran, "Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies," *Enterp Inf Syst*, vol. 14, no. 9–10, 2020, doi: 10.1080/17517575.2019.1633689.

[21] F. Xia, L. Akoglu, C. Aggarwal, and H. Liu, "Deep Anomaly Analytics: Advancing the Frontier of Anomaly Detection," 2023. doi: 10.1109/MIS.2023.3255590.

[22] S. Kumar, A. K. Kar, and P. V. Ilavarasan, "Applications of text mining in services management: A systematic literature review," 2021. doi: 10.1016/j.jjimei.2021.100008.

[23] S. Pratap, S. K. Jauhar, A. Gunasekaran, and S. S. Kamble, "Optimizing the IoT and big data embedded smart supply chains for sustainable performance," 2024. doi: 10.1016/j.cie.2023.109828.

[24] A. V. Rao and P. Shalini, "Leveraging big data and IOT for retail," *Indian J Econ Dev*, vol. 4, no. 3, 2016.

[25] K. Chaudhary and M. Alam, *Big Data Analytics: Digital Marketing and Decision-Making*. 2022. doi: 10.1201/9781003307761.

[26] K. Chaudhary and M. Alam, *Big Data Analytics; Applications in Business and Marketing*, vol. 1, no. 1. 2022.

[27] M Zaky Hadi, Ari Hasudungan Pratama Pasaribu, Fatin Saffanah Didin, and Lina Aulia, "A design of cross-selling products based on frequent itemset mining for coffee shop business," *TEKNOSAINS : Jurnal Sains, Teknologi dan Informatika*, vol. 11, no. 2, pp. 286–297, Jul. 2024, doi: 10.37373/tekno.v11i2.1065.

[28] B. Arshdeep, *Big Data Analytics: A Hands-On Approach*, vol. 21, no. 8. 2019.

[29] Y. Shi, *Advances in Big Data Analytics: Theory, Algorithms and Practices*. 2022. doi: 10.1007/978-981-16-3607-3.

[30] SHAHZAD ASLAM, "Consumer Behavior and Shopping Habits Dataset: E-Commerce Transaction Trends: A Comprehensive Dataset," https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset.

[31] G. Shmueli, P. C. Bruce, P. Gedeck, and N. R. Patel, *Data Mining for Business Analytics. Concepts, Techniques and Applications in Python*. 2020.

[32] R. Sharda, D. Delen, and E. Turban, *Business Intelligence, Analytics, and Data Science A Managerial Perspective*. 2018.

[33] M. J. Zaki and W. J. Meira, *Data Mining and Machine Learning Fundamental Concepts and Algorithms*, vol. 53, no. 9. 2020.