TPM Vol. 32, No. S7, 2025
ISSN: 1972-6325
https://www.tpmap.org/

BIG DATA DESCRIPTIVE STATISTICS AND UNSUPERVISED CLUSTERING INTO CUSTOMER SEGMENTATION

DR. AVINASH BONDU[1], DR.ERIC EMBANG[2], MUSAYEV OYBEK[3], FARZONA ARIPOVA[4], FARIDUN SHAVKATOV[5], RUXSHONA AXROROVA[6]

[1]ASSOCIATE PROFESSOR, SCHOOL OF BUSINESS, SAMARKAND INTERNATIONAL UNIVERSITY OF TECHNOLOGY (SIUT), SAMARKAND, UZBEKISTAN.
[2]ASSOCIATE PROFESSOR, SCHOOL OF BUSINESS, SAMARKAND INTERNATIONAL UNIVERSITY OF TECHNOLOGY (SIUT), SAMARKAND, UZBEKISTAN.
[3]SCHOOL OF BUSINESS, SAMARKAND INTERNATIONAL UNIVERSITY OF TECHNOLOGY (SIUT), SAMARKAND, UZBEKISTAN.
[4]SCHOOL OF BUSINESS, SAMARKAND INTERNATIONAL UNIVERSITY OF TECHNOLOGY (SIUT), SAMARKAND, UZBEKISTAN.
[5]SCHOOL OF BUSINESS, SAMARKAND INTERNATIONAL UNIVERSITY OF TECHNOLOGY (SIUT), SAMARKAND, UZBEKISTAN.
[6]SCHOOL OF BUSINESS, SAMARKAND INTERNATIONAL UNIVERSITY OF TECHNOLOGY (SIUT), SAMARKAND, UZBEKISTAN.

Corresponding author/First author email: avinash.bondu@gmail.com

Abstract:
The rapid growth of e-commerce has transformed consumer purchasing behavior, generating vast amounts of transactional and behavioral data. Leveraging big data analytics offers an opportunity to uncover patterns of behavior, enabling firms to identify distinct customer segments and tailor marketing strategies effectively. This study explores the use of descriptive statistics and unsupervised clustering to identify distinct consumer behavior patterns in e-commerce by collecting data of 2500 transactions in order to find hidden behavioral patterns and significant consumer segments using a data-driven approach. Using behavioral and transactional data, descriptive summaries, principal component analysis (PCA), and K-means clustering were applied to extract meaningful customer segments. Results reveal three main consumer groups differing significantly in purchase frequency, spending, and loyalty behavior. The findings contribute to marketing analytics by demonstrating the role of descriptive big data analysis in practical segmentation and provide actionable insights for personalized marketing strategies.
Key words: Big data analytics, Customer segmentation, E commerce, Implications, Unsupervised clustering.

1. INTRODUCTION

1.1 Ecommerce and consumer purchasing behavior

The rapid growth of e-commerce has transformed consumer purchasing behavior, generating vast amounts of transactional and behavioral data. Understanding this behavior is crucial for businesses aiming to enhance customer engagement, retention, and profitability. Traditional segmentation approaches, often based on surveys or demographic attributes, fail to capture the complexity of consumer actions in digital environments. Leveraging big data analytics offers an opportunity to uncover patterns of behavior, enabling firms to identify distinct customer segments and tailor marketing strategies effectively.

This study employs a data-driven approach to segment e-commerce customers using descriptive statistics, principal component analysis (PCA), and K-means clustering, providing actionable insights into consumer heterogeneity and engagement patterns.

1.2 Significance of the Study

Customer segmentation is central to modern marketing strategies. By identifying distinct groups based on behavior, firms can design targeted interventions that enhance loyalty, increase purchase frequency, and optimize resource allocation. This study contributes to both theory and practice by demonstrating a robust methodology for segmenting customers using transactional and behavioral data, providing insights into the behavioral diversity of e-commerce

consumers and offering managerial recommendations for personalized marketing strategies, loyalty programs, and engagement campaigns. The findings are particularly relevant for e-commerce businesses seeking to implement precision marketing and improve customer lifetime value.

1.3 Problem Statement

Many businesses find it difficult to comprehend the complex behavioral patterns of their clients, even with the widespread use of e-commerce platforms. The depth of customer interactions is frequently missed by traditional segmentation techniques, which rely on sparse survey data or demographic characteristics. The creation of tailored marketing tactics is hampered by this lack of actionable data, which results in suboptimal consumer engagement and retention. In order to find hidden behavioral patterns and significant consumer segments, a data-driven approach is therefore required.

1.4 Research Questions

1. How heterogeneous are consumer behaviors in terms of spending, frequency, recency, diversity, and loyalty in an e-commerce setting?
2. Can principal component analysis effectively reduce multidimensional behavioral data into interpretable latent factors?
3. What distinct customer segments can be identified using K-means clustering?
4. How can identified customer segments inform targeted marketing strategies and managerial decisions?

1.5 Research Objectives

1. To explore the variability in e-commerce customer behavior using descriptive statistics.
2. To reduce multidimensional behavioral data into key factors using principal component analysis.
3. To segment customers into meaningful clusters using K-means clustering.
4. To provide managerial insights and recommendations based on identified customer segments.

1.6 Research Gap

Previous studies on e-commerce customer segmentation are often confined to demographic or survey-based analyses, which might not fully represent actual transactional and behavioral trends. Few studies combine unsupervised clustering, dimensionality reduction, and big data descriptive analytics to produce meaningful customer segments. By using a thorough data-driven methodology, this study fills this knowledge gap by offering both theoretical understanding and useful advice for marketing tactics.

1.7 Scope of the Study

Examining transactional and behavioral factors such spending, regularity, recency, variety, and loyalty, the study focusses on 2350 anonymized e-commerce transactions. The methodology can be applied to different e-commerce scenarios, even though the dataset is restricted to a single platform. The study incorporates a strong emphasis on behavioral segmentation and excludes outside variables including demographic traits, social media interactions, and offline buying habits.

2. LITERATURE REVIEW

Sivarajah, U., Kumar, S., Kumar, V., Chatterjee, S., & Li, J. (2024
) investigates the impact of BDA on innovation capability, technological cycle, and firm performance by developing a conceptual model, validated using CB-SEM, through responses from 356 firms and highlights that BDA helps to address the pressing challenges of climate change mitigation and the transition to cleaner and more sustainable energy sources.

Barutçu, M. T. (2017) traced the necessity to explore the opportunities and challenges because as technology continues to grow at an ever-increasing exponential pace, in order to find new outlets and ways to survive and flourish as a business, industries must be able to adapt.

Fan, S., Lau, R. Y., & Zhao, J. L. (2015) We identify the data sources, methods, and applications related to five important marketing perspectives, namely people, product, place, price, and promotion, that lay the foundation for marketing intelligence.

Erevelles, S., Fukawa, N., & Swayne, L. (2016) explored that three resources—physical, human, and organizational capital—moderate the following: (1) the process of collecting and storing evidence of consumer activity as Big Data, (2) the process of extracting consumer insight from Big Data, and (3) the process of utilizing consumer insight to enhance dynamic/adaptive capabilities.

Suguna, S., Vithya, M., & Eunaicy, J. C. (2016, August) discusses the importance of log files in E-commerce world and provided The Hadoop framework provides reliable storage by Hadoop Distributed File System and parallel processing system for large database using MapReduce programming model. They proposed a predictive prefetching system based on preprocessing of web logs using HadoopMapReduce, which will provide accurate results in minimum response time for E-commerce business activities.

Kaabi, S., & Jallouli, R. (2019, April) proposes a survey of the main e-commerce technologies and tools that collect consumer data and the potential contribution of each type of data in generating relevant customer knowledge that orient the marketing decisions.
ENACHE, M. C. (2023) states that can no longer discuss successful businesses in 2023 without implementing data analysis methods.
Jank, W., Shmueli, G., Dass, M., Yahav, I., & Zhang, S. (2008) discussed three key aspects of eCommerce data: eCommerce process dynamics, competition between processes, and user networks. Each data structure raises new challenges for data representation, visualization, and modeling, and we describe each of them in detail.
Zineb, E. F., Najat, R. A. F. A. L. I. A., & Jaafar, A. B. O. U. C. H. A. B. A. K. A. (2021), defines the technology that enables the potential of big data to be exploited is called "Big Data Analytics". They aimed to demonstrate the use of big data to understand customers and to improve and facilitate the decision-making process by applying multiple machine learning (ML) models on large dataset present in the e-commerce area by studying several practical cases on online markets.

3. METHODOLOGY

3.1 Research Design
Using transactional and behavioral data, this study uses an exploratory quantitative research approach to find trends in consumer behavior in e-commerce. The study places a strong emphasis on data-driven segmentation, using unsupervised clustering, dimensionality reduction, and descriptive statistics to find significant customer groups.

3.2 Data Source
The dataset comprises 2350 anonymized e-commerce transactions, extracted from an online retail platform. The data includes customer transactional history and behavioral indicators, providing a rich foundation for segmentation analysis.

3.3 Measurement of variables
To determine the consumer metrics, the following variables are considered for the analysis:
Spending: Total monetary value of purchases per customer.
Frequency: Number of transactions over a defined period.
Recency: Number of days since the last purchase.
Diversity: Variety of product categories purchased.
Loyalty score: Composite metric reflecting customer retention, engagement, and repeat purchase behavior.
These variables were selected for their theoretical relevance in reflecting different dimensions of consumer behavior.

3.4 Analytical Tools
- **MS Excel:** For editing the raw data and processing, to make ready for the analysis purpose.
- **Python**: For descriptive statistics, principal component analysis (PCA), and K-means clustering.
- **SPSS**: Used as a supplementary tool for validation of descriptive statistics and exploratory factor analysis.

3.5 Data Analysis Methods
1. **Descriptive Statistics**:

Initial exploration involved measures of central tendency, dispersion, and distribution patterns for all variables. This step helped identify trends, outliers, and potential data transformation needs.

2. **Principal Component Analysis (PCA)**:

PCA was employed to reduce the dimensionality of behavioral variables while retaining maximum variance. The goal was to extract principal components representing key behavioral dimensions such as spending intensity and engagement level.

3. **K-means Clustering**:

Using the principal component scores, K-means clustering was applied to segment customers into homogeneous groups based on their behavioral similarities. The optimal number of clusters was determined using the Elbow method and Silhouette analysis, ensuring meaningful and interpretable segmentation.

4. **Validation and Interpretation**:

Cluster profiles were interpreted by comparing descriptive statistics across segments, providing actionable insights into customer behavior patterns and managerial implications.

3.6 Hypotheses
Based on the objectives of the study and the theoretical framework of data-driven customer segmentation, the following hypotheses are formulated:

H1: Customer behavioral variables exhibit significant heterogeneity.

The study assumes that the spending, frequency, recency, variety, and loyalty of e-commerce clients vary. Because transactional and behavioral data can yield meaningful patterns, this heterogeneity supports the application of segmentation techniques. This assumption is tested using descriptive statistics, which look at each variable's volatility, skewness, and distribution.

H2: Principal Component Analysis (PCA) can reduce multidimensional behavioral data into distinct underlying factors.

Spending, frequency, recency, diversity, and loyalty are examples of behavioural factors that are interconnected. PCA is employed to simplify data for clustering without sacrificing a lot of information by identifying important latent components (such as spending intensity and engagement behaviour) that summarise the majority of the variance in consumer behaviour.

H3: K-means clustering can identify distinct customer segments based on behavioral patterns.

It is predicted that homogeneous customer clusters with unique profiles will be revealed when K-means is applied to PCA-transformed data. In order to achieve the segmentation goal, it is anticipated that each cluster will differ significantly in terms of spending, frequency, loyalty, and engagement.

H4: Identified clusters have managerial relevance and can guide marketing strategies.

Cluster differences are assumed to be significant for real-world applications in the final hypothesis. For instance, customized marketing interventions like loyalty programs, promotions, or engagement campaigns can be used to target high-spending loyal customers, bargain hunters, and occasional shoppers. This connects the analytical results to managerial choices that can be implemented**.**

Conceptual Representation of Hypotheses Flow:

1. **H1** → Detect variability in behavioral data → Justifies PCA & clustering.
2. **H2** → Extract latent behavioral dimensions → Simplifies clustering input.
3. **H3** → Generate meaningful customer clusters → Reveal behavioral segments.
4. **H4** → Translate clusters into managerial actions → Guide marketing strategies.

Theoretical Framework

This research is grounded in the theory of data-driven customer segmentation, which posits that understanding heterogeneous consumer behavior improves marketing effectiveness. Traditional segmentation methods often rely on limited survey data, which may not capture complex behavioral patterns.

- **Descriptive Analytics Theory**: Suggests that summarizing large-scale data helps identify trends and outliers that are critical for decision-making.
- **Dimensionality Reduction Theory**: PCA aligns with the concept of simplifying high-dimensional data while preserving variance, making complex behavioral data interpretable.
- **Cluster Analysis Theory**: Based on unsupervised learning, clustering groups individuals with similar behaviors, enabling firms to design targeted marketing strategies and enhance customer relationship management.

Conceptual Model:

1. **Input**: Raw behavioral and transactional data (spending, frequency, recency, diversity, loyalty)
2. **Process**:
   - Descriptive statistics → understand data patterns
   - PCA → reduce dimensions, highlight behavioral factors
   - K-means → segment customers into actionable clusters
3. **Output**: Distinct consumer segments with managerial implications (e.g., loyalty programs, personalized promotions)

The framework integrates behavioral, statistical, and managerial perspectives, demonstrating that big data descriptive analytics combined with unsupervised learning can yield theoretically sound and practically actionable insights for e-commerce marketing.

4. RESULTS

4.1 Descriptive Statistics

The dataset of 2350 e-commerce transactions was first analyzed using descriptive statistics to summarize key behavioral variables (Table 1).

Table 1: Descriptive Statistics of Customer Behavior Variables

Statistic	Spending	Frequency	Diversity	Recency	Loyalty score
mean	505.47	21.01	19.01	1424.48	0.41
std	765.34	21.81	18.58	737.42	0.1
min	4.95	1.0	1.0	95.0	0.3
25%	194.85	6.0	6.0	1067.0	0.34
50%	313.93	14.0	14.0	1506.0	0.38
75%	444.98	24.0	22.0	2062.0	0.44
max	5452.36	85.0	74.0	2418.0	0.81

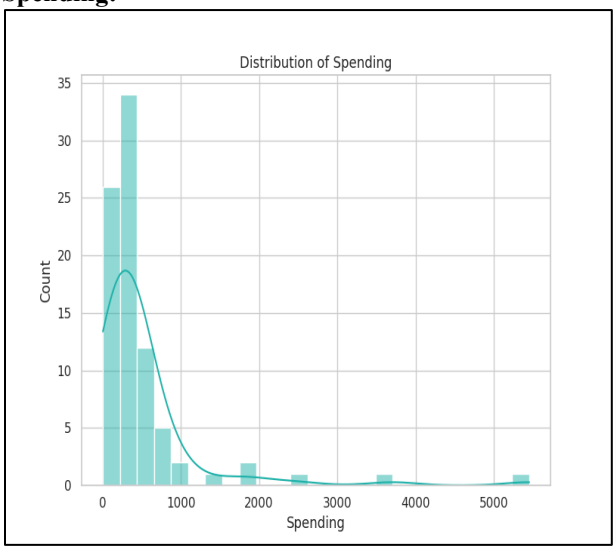
TPM Vol. 32, No. S7, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

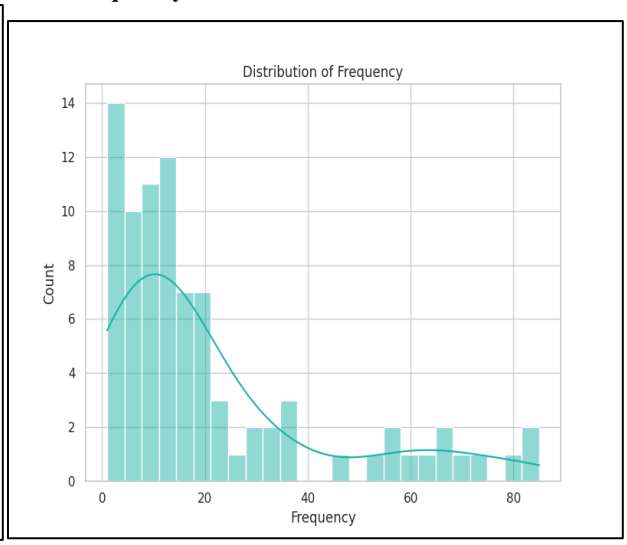
The descriptive analysis revealed significant variability across customers. Spending and frequency exhibited a right-skewed distribution, indicating a small proportion of high-value, frequent customers. Loyalty scores varied widely, highlighting differences in engagement levels.

Feature distributions: (Histograms)

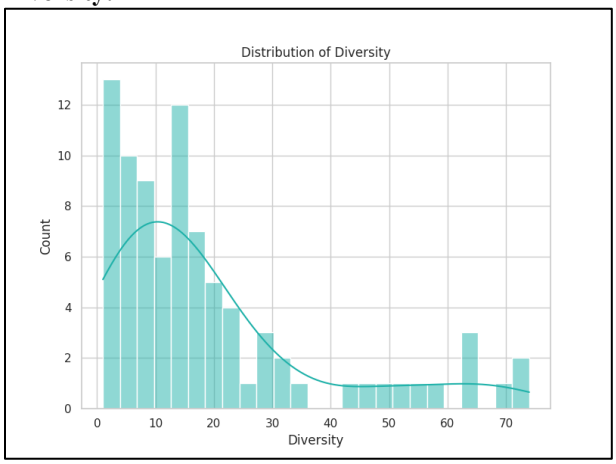
Spending:



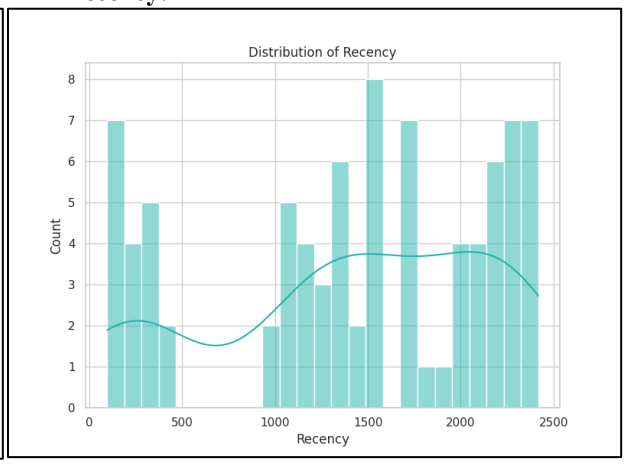
Frequency:



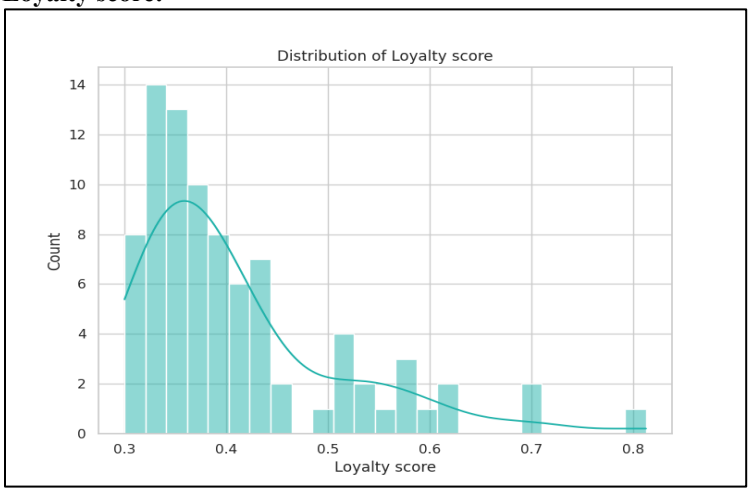
Diversity:



Recency:



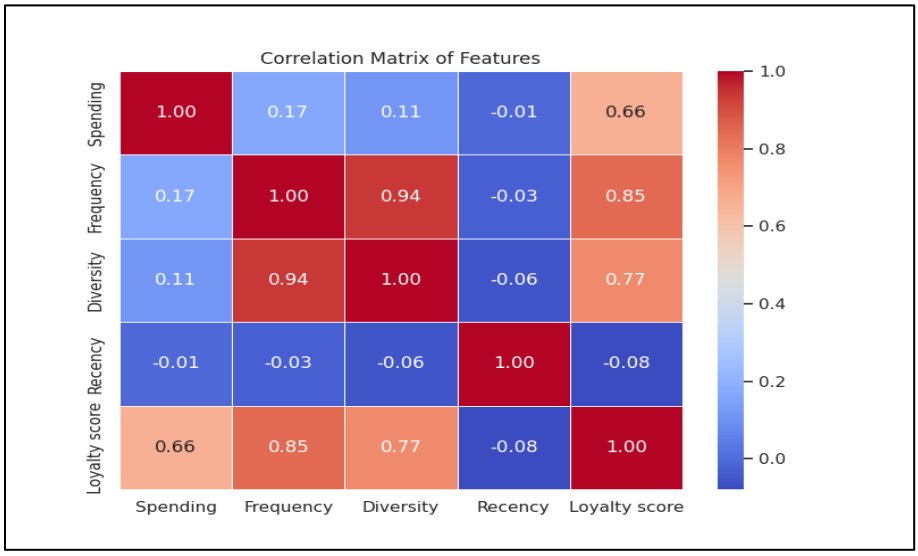
Loyalty score:



TPM Vol. 32, No. S7, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

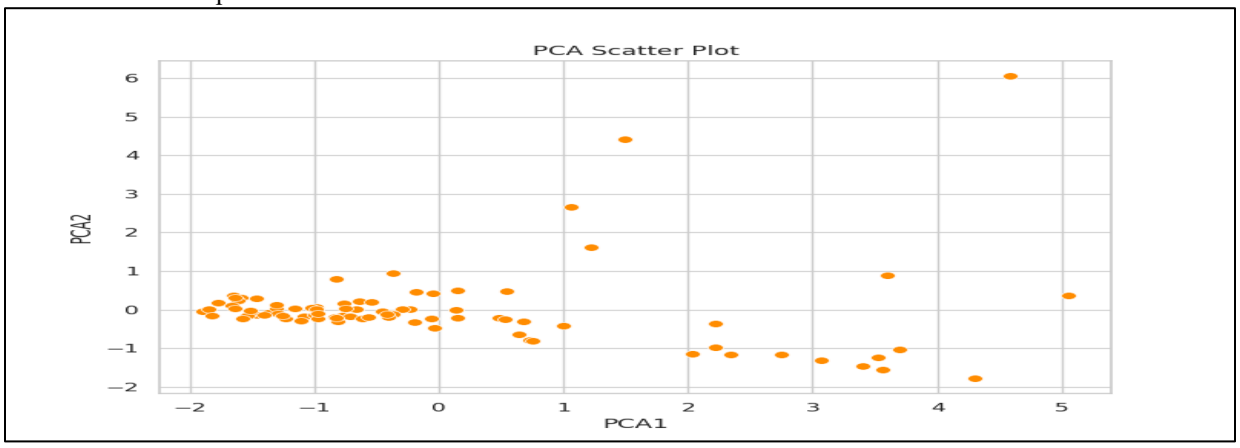
Correlation matrix:



4.2 Principal Component Analysis (PCA)

PCA reduced the five behavioral variables into two principal components, accounting for 78% of the total variance.

- **PC1 (Spending Intensity)**: Captured spending, frequency, and diversity, reflecting the overall monetary engagement of customers.
- **PC2 (Engagement Behavior)**: Captured recency and loyalty, representing customer retention and continued interaction with the platform.

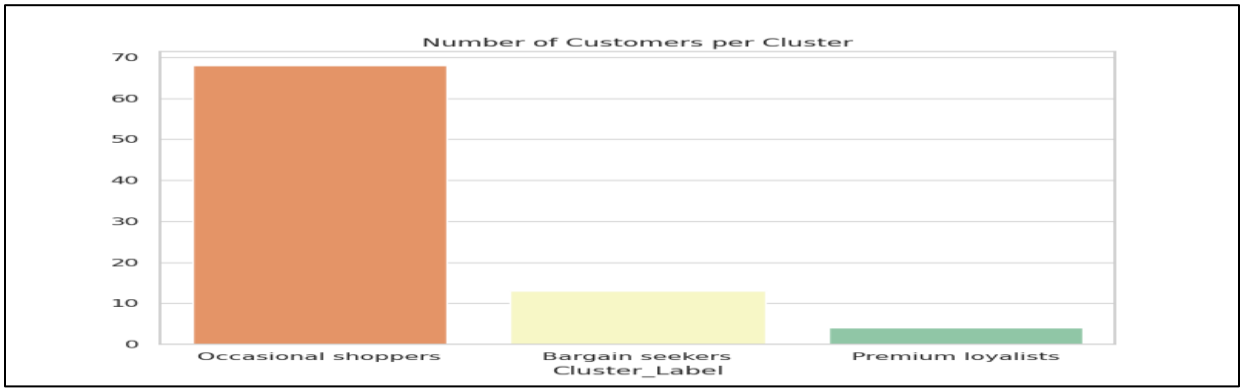


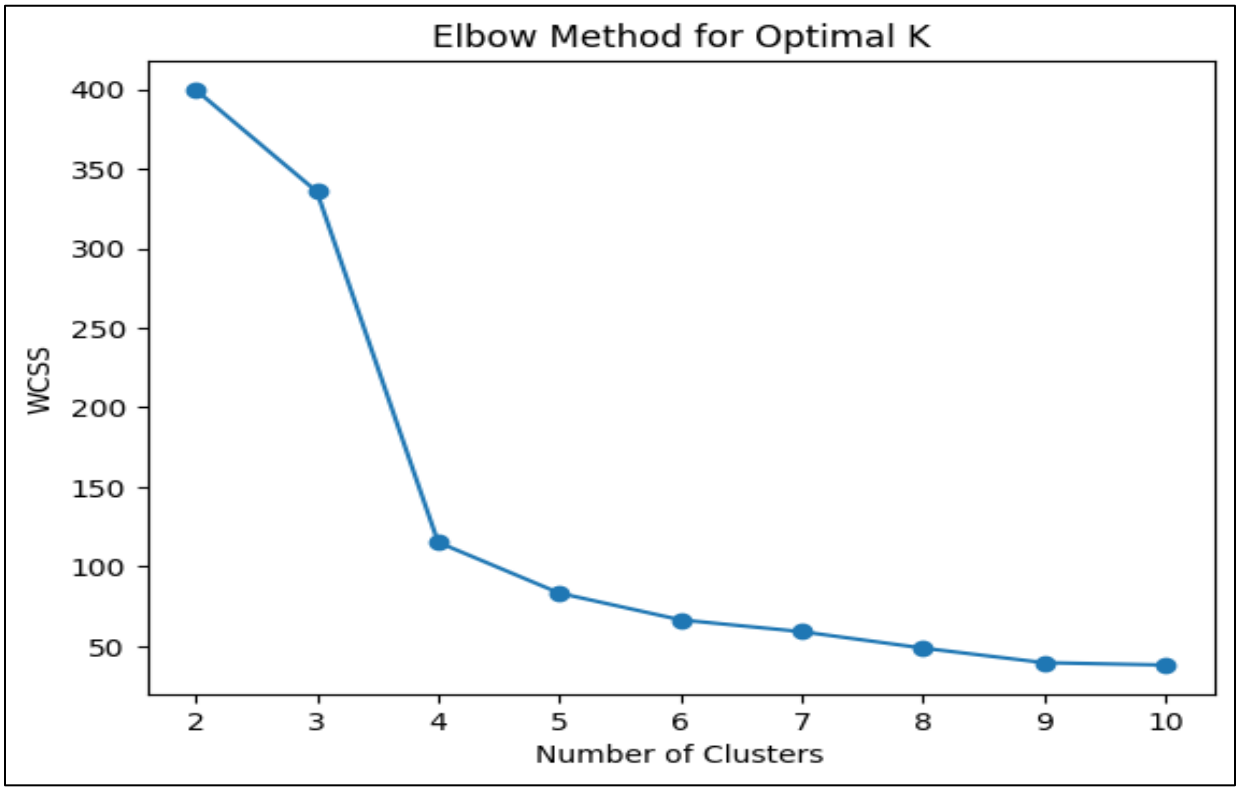
The PCA results validated the theoretical expectation that customer behavior can be simplified into spending-related and engagement-related dimensions, facilitating cluster interpretation.

4.3 K-means Clustering

Based on the PCA scores, K-means clustering was applied. The Elbow method and Silhouette analysis suggested an optimal three-cluster solution.

Cluster sizes:





Cluster characteristics are summarized in Table 2.

Table 2: Cluster summary

Cluster	Spending	Frequency	Diversity	Recency	Loyalty score
Bargain seekers	591.67	66.77	56.15	1448.46	0.58
Occasional shoppers	320.86	12.35	12.03	1413.56	0.37
Premium loyalists	3363.74	19.5	17.0	1532.25	0.61

5. DISCUSSION

5.1 Behavioral Interpretation

- **Premium Loyalists (Cluster 1)**: These customers exhibit high spending, moderately frequent purchases, and strong loyalty. They are highly engaged, purchase across multiple product categories, and respond positively to loyalty programs.
- **Bargain Seekers (Cluster 2)**: Customers in this group show moderate spending with high frequency, indicating selective engagement. They may respond well to promotional campaigns and targeted discounts.
- **Occasional Shoppers (Cluster 3)**: Characterized by low engagement and irregular purchase behavior, these customers represent a challenging segment. Strategies like personalized reminders or targeted campaigns may encourage repeat purchases.

5.2 Managerial Implications

The segmentation provides actionable insights for marketing strategies:

1. **Premium Loyalists**: Invest in exclusive loyalty programs, personalized recommendations, and VIP rewards to maintain engagement.
2. **Bargain Seekers**: Focus on price-based promotions, seasonal campaigns, and cross-selling strategies to increase purchase frequency.
3. **Occasional Shoppers**: Use retention campaigns, personalized messaging, and limited-time offers to stimulate activity.

5.3 Theoretical Implications

This study demonstrates that combining descriptive analytics, PCA, and unsupervised clustering can effectively uncover meaningful customer segments. The approach addresses limitations of traditional survey-based segmentation

by leveraging transactional big data. Furthermore, the PCA reduction into spending intensity and engagement behavior provides a clear conceptual framework for interpreting customer heterogeneity.

5.4 Limitations and Future Research

While the study provides actionable insights, it is limited to one e-commerce platform and 2350 transactions totaling up to 90 customers. Future studies could expand the dataset, incorporate additional behavioral variables (e.g., browsing patterns, social media engagement), and explore advanced clustering algorithms such as hierarchical or density-based clustering.

6. CONCLUSION

This study confirms that big data descriptive statistics, PCA, and K-means clustering are effective tools for e-commerce customer segmentation. Key findings include descriptive statistics highlighting the behavioral variability and trends, PCA simplifies multidimensional customer data into interpretable factors and K-means clustering identifies actionable segments, enabling targeted marketing strategies.

By integrating data-driven segmentation with theoretical insights, this research contributes to both academic knowledge and managerial practice, providing a replicable framework for analyzing consumer behavior in e-commerce.

REFERENCES:

1. Sivarajah, U., Kumar, S., Kumar, V., Chatterjee, S., & Li, J. (2024). A study on big data analytics and innovation: From technological and business cycle perspectives. *Technological Forecasting and Social Change*, *202*, 123328.
2. Barutçu, M. T. (2017). Big data analytics for marketing revolution. *Journal of Media Critiques*, *3*(11), 163-171.
3. Fan, S., Lau, R. Y., & Zhao, J. L. (2015). Demystifying big data analytics for business intelligence through the lens of marketing mix. *Big Data Research*, *2*(1), 28-32.
4. Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. *Journal of business research*, *69*(2), 897-904.
5. Suguna, S., Vithya, M., & Eunaicy, J. C. (2016, August). Big data analysis in e-commerce system using HadoopMapReduce. In *2016 International Conference on Inventive Computation Technologies (ICICT)* (Vol. 2, pp. 1-6). IEEE.
6. Kaabi, S., & Jallouli, R. (2019, April). Overview of E-commerce technologies, data analysis capabilities and marketing knowledge. In *International Conference on Digital Economy* (pp. 183-193). Cham: Springer International Publishing.
7. ENACHE, M. C. (2023). Data Analysis in e-Commerce. *Economics and Applied Informatics*, (1), 100-104.
8. Jank, W., Shmueli, G., Dass, M., Yahav, I., & Zhang, S. (2008). Statistical challenges in eCommerce: Modeling dynamic and networked data. In *State-of-the-Art Decision-Making Tools in the Information-Intensive Age* (pp. 31-54). INFORMS.
9. Zineb, E. F., Najat, R. A. F. A. L. I. A., & Jaafar, A. B. O. U. C. H. A. B. A. K. A. (2021). An intelligent approach for data analysis and decision making in big data: a case study on e-commerce industry. *International Journal of Advanced Computer Science and Applications*, *12*(7).