IJIRR

International Journal of Information Research and Review
Vol. 12, Issue, 04, pp.7997-8002, April, 2025

RESEARCH ARTICLE

CLICKS AND SHOPPERS: IDENTIFYING CONSUMER PATTERNS

Dr. Rajesh A., Akshaya, S., Pradheena, S. and Santhiya, M.

Department Of Information Technology, Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, An Autonomous Institution Affiliated By Anna University, Chennai

ARTICLE INFO

Article History:
Received 18th January, 2025
Received in revised form
24th February, 2025
Accepted 25th March, 2025
Published online 23rd April, 2025

Key words:
Clothing Retail, Big Data Analytics, Clickstream Analysis, Customer Segmentation, Online Purchasing.

ABSTRACT

This study analyzes consumer segments within the online retail sector using the Online Shoppers Intention Dataset. Six unique segments were identified: Regular Shoppers, the largest segment, consisting mainly of returning visitors with low spending; General Visitors, who show increased activity in the second half of the year; End-Year Shoppers, primarily active in the last quarter; Non-Buyers, with the highest exit rates and minimal revenue; Holiday Shoppers, who predominantly make purchases in November and December; and Frequent Buyers, a smaller group with the highest purchase rate. These insights can help e-retailers better target specific segments and boost revenue generation.

INTRODUCTION

The rise of fast fashion has dramatically reshaped the retail industry, focusing on rapid production and distribution of trendy, affordable products. This model has played a significant role in making products widely accessible and popular across global markets. Brands that prioritize quick design-to-shelf cycles have experienced exceptional growth over the years. Consequently, these brands have become some of the most influential players in the retail sector, similar to major tech companies like Apple and Google. This widespread success necessitates a deeper understanding of the factors that drive these brands' prominence, particularly in relation to evolving consumer behaviors and the role of technology in modern retail. A primary factor behind the success of fast fashion has been the changing preferences and behaviors of consumers. Today's shoppers are more value- conscious, seeking a continuous stream of new products at affordable prices. This shift has created an insatiable demand for goods that cater to a wide variety of tastes, often with a fast turnover rate. Additionally, advancements in digital technologies and the growth of e-commerce have significantly altered the retail experience. The increasing preference for online shopping has prompted many fast-fashion brands to strengthen their digital presence, using e-commerce platforms not only to drive sales but also as a tool to collect customer insights and refine business strategies. While many fast-fashion brands have benefited from this digital transformation, not all have fully capitalized on the opportunities that come with it. The most successful brands are those that have leveraged data analytics to make informed decisions. By utilizing large volumes of data, these companies can optimize inventory management, marketing, and customer targeting. For example, analyzing customer interactions on e-commerce platforms, such as clicks, page visits, and purchasing behavior, has provided brands with valuable insights into customer preferences. This data helps brands personalize the shopping experience and optimize product recommendations.Despite the promise of clickstream data for marketing and consumer analysis, many retailers have not fully tapped into its potential. Clickstream data, which tracks consumers' navigation on websites, offers a wealth of information on browsing habits, interests, and purchase intentions. However, the vast volume and complexity of this data present challenges for many brands, as it requires sophisticated tools to process. While a few leading brands have begun integrating machine learning algorithms to analyze clickstream data, the broader industry is still catching up in terms of using this information to shape marketing strategies.

One area where clickstream data can have a profound impact is in consumer segmentation. Traditionally, segmentation has relied on demographic or psychographic data, but with the growth of online retail, there is an increasing emphasis on segmenting consumers based on their digital behavior. While research on offline consumer segmentation is abundant, there is a gap in understanding online consumer segments—especially in terms of identifying which segments drive the most revenue. This study aims to address this gap by focusing on an advanced approach to consumer segmentation that links online behavior with revenue generation potential. Through unsupervised machine learning techniques, the research seeks

to identify distinct consumer segments based on their online activities—such as page views, browsing patterns, and session length—and correlate these segments with their likely revenue contribution. Understanding which consumer segments are most likely to convert into high-value customers is crucial for optimizing marketing efforts and maximizing return on investment in digital advertising and promotions. By segmenting consumers based on their revenue potential, retailers can target their efforts more effectively, focusing resources on the segments most likely to yield the highest returns. This approach also allows companies to tailor their product offerings and consumer experiences to better meet the needs of their most profitable consumers, which in turn improves customer satisfaction and retention. This study applies unsupervised clustering algorithms to analyze consumer behavior using clickstream data. The findings contribute to the growing body of knowledge in digital marketing and e-commerce analytics, showing how clustering techniques can help identify high-value consumer segments. Additionally, the research sheds light on the broader role of big data in shaping modern retail strategies, offering a fresh perspective on how digital behavior can be leveraged to drive business success in the online retail environment.

Ultimately, the aim of this research is to highlight the importance of data-driven decision-making in modern retail and offer a framework for using consumer segmentation as a tool to enhance profitability in e-commerce. By doing so, this study not only addresses an important gap in the existing literature but also provides practical insights for retailers seeking to leverage big data to stay ahead of the competition in an increasingly digital world.

LITERATURE SURVEY

The fast fashion industry has seen rapid growth due to its ability to adapt quickly to shifting consumer preferences. Factors like seasonality, trends, and social media significantly influence these preferences, and as consumer behavior becomes more fragmented and digital, retailers are challenged to remain relevant. This makes it essential for businesses to use data-driven strategies that help them understand and respond to consumer behavior patterns. One of the key tools in achieving this is clickstream data analysis. This involves tracking a consumer's online behavior—how they browse and interact with a website—to gain insights into their preferences and purchasing decisions. Studies by Yu et al. (2018) and Schellong et al. (2016) show how clickstream data allows businesses to identify distinct consumer segments, predict future purchasing patterns, and refine the online shopping experience. By examining how users navigate a website, businesses can determine whether they are simply browsing for inspiration, actively searching for products, or ready to make a purchase. This segmentation of consumers is invaluable for understanding how shoppers react to trends, promotions, and how they move through the sales funnel. Advanced machine learning algorithms further enhance these insights by predicting long-term consumer behavior and optimizing inventory management (Chamberlain et al., 2017). Social media data also plays a transformative role in how fashion brands interact with consumers. Caro and Martinez-de-Albeniz (2015) highlight that social media platforms are vital for brands to engage with their audience, promote products, and influence purchasing decisions. Brands leverage social media by encouraging user-generated content and influencer partnerships to create marketing campaigns that align with consumer values, preferences, and aspirations. This is further amplified by an omnichannel approach, which ensures consistent brand presence across both digital and physical platforms, enabling brands to reach a wider, more diverse customer base (Mosquera et al., 2019). Sustainability and ethical fashion are additional trends gaining prominence in the industry. With growing awareness of environmental issues, more consumers are gravitating toward brands that prioritize sustainability, ethical labor practices, and eco-friendly materials. Choi et al. (2017) argue that integrating sustainability into the core business model not only aligns with consumer values but can also provide a competitive advantage. Data analytics can help brands track and communicate their sustainability efforts, thereby increasing consumer trust and loyalty. As consumers demand greater transparency, digital platforms can hold brands accountable, encouraging them to adopt more sustainable practices.

Big data analytics has revolutionized how fashion brands manage their supply chains and forecast demand. By providing detailed insights into consumer preferences and behaviors, big data enables businesses to make informed decisions about product development, inventory management, and pricing strategies. Chan et al. (2017) note that predictive analytics, derived from big data, allows businesses to anticipate future trends, enabling them to not only react to current consumer demands but also plan for long-term success. This shift in data-driven decision-making has helped brands minimize waste, optimize production, and reduce costs.

However, applying big data analytics comes with its challenges. Issues surrounding data privacy, ethical data usage, and the integration of disparate data sources remain prevalent (Choi et al., 2019). Smaller brands, in particular, may struggle to keep up with the digital transformation due to resource constraints, making it necessary to explore ways for small and medium-sized enterprises (SMEs) to leverage data analytics without overwhelming their resources. Furthermore, as consumer data is increasingly used to segment and target customers, ethical concerns regarding data collection, the use of artificial intelligence in decision-making, and algorithmic biases have emerged. These concerns need to be addressed responsibly to ensure that technologies like AI are used fairly and transparently.

Clickstream data offers substantial insights into consumer intentions and purchase behavior. By examining the consumer journey—from discovery to final purchase—fashion retailers can uncover real-time data about what drives purchasing decisions. However, the raw clickstream data needs to be processed using sophisticated tools and algorithms to extract meaningful patterns and trends. This analysis enables retailers to segment customers effectively, grouping them based on their browsing behavior. Techniques like k-means clustering are commonly used for segmentation, but newer algorithms such as partitioning around medoids (PAM) have become more popular for their ability to handle noise and provide more accurate segmentation (Gupta and Sharma, 2022).

Predictive analytics is changing the way fashion retailers approach customer segmentation. By applying machine learning to historical clickstream data, retailers can predict future purchasing behavior and emerging trends. This allows for better stock management, pricing decisions, and product

recommendations. The adoption of AI and machine learning continues to reshape how brands analyze consumer data and engage with customers. However, these technologies require substantial investment, skilled personnel, and specialized tools—resources that might be challenging for smaller companies to afford. Data privacy concerns have also become more pronounced as consumers become increasingly aware of how their personal data is used. Retailers must strike a balance between leveraging data for insights and adhering to ethical data practices, ensuring compliance with regulations like the General Data Protection Regulation (GDPR). These considerations are essential for maintaining consumer trust, which is crucial for long-term success.

In conclusion, data plays an essential role in the fast fashion industry, enabling businesses to respond to consumer demands swiftly and effectively. By using clickstream data and advanced analytics, brands can identify consumer segments, predict trends, and refine their offerings to stay competitive. However, ethical challenges surrounding data collection and usage must be addressed carefully. As the industry evolves, the ability to innovate and adapt to data-driven strategies will determine the success of fast fashion brands.

METHODOLOGY

This study uses a comprehensive methodology for identifying distinct consumer behavior segments in a fast-fashion retailer's clickstream dataset. The dataset is highly complex, comprising continuous, categorical, and binary variables, which necessitates a multi-step approach. These steps include data preparation, feature selection, dimensionality reduction, clustering analysis, and cluster validation. K-means clustering was chosen as the primary technique due to its efficiency in handling large datasets, while Partitioning Around Medoids (PAM) with Gower distance was considered as a secondary method, especially for mixed data types.

Data Preparation and Cleaning: The initial data preparation phase involves handling missing values and outliers to ensure the accuracy of clustering results. Given that the dataset includes a mix of numerical and categorical features, preprocessing was essential to normalize the data and mitigate the effects of noise and incomplete entries.

- **Handling Missing Values:** For continuous variables with missing data, the most common approach is to impute missing values with the mean or median of the feature. Imputation is performed to maintain the integrity of the dataset without distorting its structure.
- **Outlier Removal:** Outliers were identified using the standard deviation method. Any data point more than three standard deviations away from the mean was considered an outlier. The formula for outlier detection is:

Outlier Threshold = $\mu \pm 3\sigma$

Where μ is the mean and σ is the standard deviation of the feature. Once identified, these outliers were either adjusted or removed to avoid skewing the results during clustering.

Feature Selection: To improve the performance of clustering algorithms, feature selection was performed to reduce the dimensionality of the dataset. Feature importance was calculated using a Random Forest algorithm, which provides a score indicating the importance of each feature. The importance of a feature iii was calculated by averaging its importance across multiple decision trees in the forest:

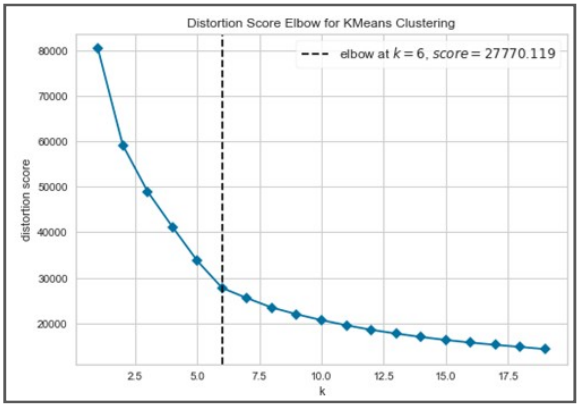
$$\text{Importance}(i) = \frac{1}{N}\sum_{j=1}^{N} \text{imp}_j(i)$$

Dimensionality Reduction with PCA: Principal Component Analysis (PCA) was applied to reduce the number of features while preserving as much variance as possible. PCA is particularly useful when dealing with high-dimensional data, as it simplifies the dataset into principal components that explain the majority of the variance. The variance explained by each principal component is given by:

Variance Explained = λ_i / Σ (λ_j)

from j = 1 to p, where λ_i is the eigenvalue of the i-th principal component, and p is the total number of components. In this study, the first few principal components that explained over 85% of the variance were retained, significantly reducing the dimensionality of the data and aiding in more efficient clustering

Clustering with K-means: K-means clustering was chosen for its efficiency and scalability, particularly when dealing with large datasets like the one at hand. The algorithm works by partitioning the data into K clusters, with each data point assigned to the cluster whose center is the closest. The objective function for K-means clustering, which minimizes the sum of squared Euclidean distances between points and their cluster centers. The elbow method was used to determine the optimal number of clusters by plotting the variance explained as a function of K and identifying the "elbow" where the increase in explained variance diminishes.



This Elbow Plot helps determine the optimal number of clusters (k) for K-means clustering. The "elbow" at k = 6 indicates the best choice, as adding more clusters beyond this point yields minimal improvement in reducing the distortion score (around 27770.119 at k=6). Thus, 6 clusters is the most efficient choice for this dataset.

Exploration of PAM with Gower Distance: An alternative to K-means, Partitioning Around Medoids (PAM) was tested using Gower distance, which is particularly suitable for datasets containing both continuous and categorical variables. Gower distance is calculated as the average of distances for each feature between two points xand y.

Despite its robustness for mixed data types, K-means clustering proved to be more computationally efficient after dimensionality reduction through PCA.

Cluster Validation: Once the clusters were formed, their validity was evaluated using the silhouette score, which provides an indication of how well each data point fits into its assigned cluster. The silhouette score for a data point iii is calculated as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Where a(i) is the average distance from point i to all other points in the same cluster, and b(i) is the average distance from point i to all points in the nearest cluster. A silhouette score closer to 1 indicates that the point is well-clustered, while a score closer to -1 indicates that the point is poorly clustered. An average silhouette score above 0.5 was considered indicative of well-separated and well-formed clusters.

Insights into Segment Profitability: Following clustering, the profitability of each segment was evaluated using the Kruskal-Wallis test, a non- parametric method used to compare differences between more than two groups. The Kruskal-Wallis statistic is given by:

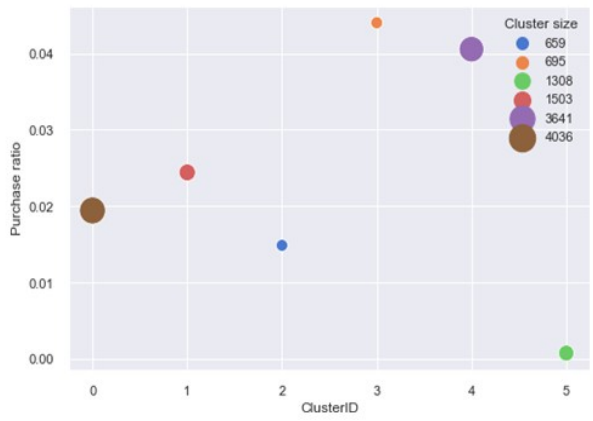
$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\bar{R}_i - \bar{R}\right)^2$$

Where ni is the number of observations in the ith group, $\bar{R}_i$ is the average rank of the ith group $\bar{R}$ is the overall average rank. This test helped identify statistically significant differences in profitability between the segments, further informing the retailer's marketing strategy.

Visualizations and Data Interpretation: To aid in the interpretation of clustering results, various visualizations were created. The elbow method was plotted to determine the optimal number of clusters, and silhouette score charts were used to assess cluster validity. PCA scatter plots helped visualize the distribution of data in reduced dimensions, revealing the distinct separation between clusters. Additionally, box plots and heatmaps were employed to illustrate the distribution of key features within each segment, providing further insights into consumer behavior.

RESULTS

Cluster Analysis



In this analysis, we performed clustering on the dataset and identified six distinct clusters based on visitor behaviors. Here's a summary of each cluster:

Cluster 0: This is the largest cluster, containing 90% of returning visitors. Visitors in this group are most likely to shop during May. The high proportion of returning visitors indicates a strong preference for repeat visits, suggesting that this group may be highly loyal to the platform.

Cluster 1: This cluster seems to represent more general visitors. These visitors are more likely to make purchases in the second half of the year. One of the key characteristics of this cluster is that visitors spend the most time on administrative pages, which may imply that they are exploring detailed product specifications, policies, or account-related information before making a purchase.

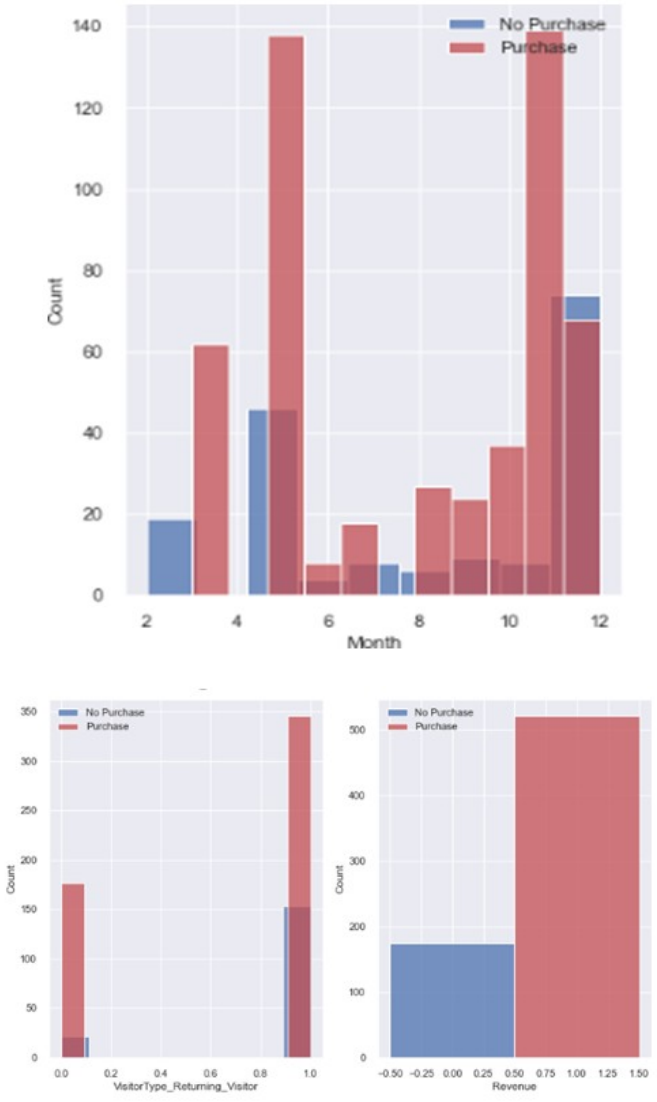
Cluster 2: Cluster 2 has characteristics similar to Cluster 1 but with a notable difference. Visitors in this cluster are more likely to shop during the last quarter of the year. They spend more time on Information and Product-related pages, indicating a stronger interest in product details and company information before making a purchase.

Cluster 3: This cluster has the highest purchase ratio, though it is only the second-largest in size. Visitors in this cluster are typically returning visitors who spend significantly less time on administrative and informational pages of the website. Their quick purchase decisions and high purchase ratio suggest they are more focused on the buying process rather than browsing product details or company policies.

Cluster 4: Cluster 4 is of particular interest due to its second-highest purchase ratio and size**.** Visitors in this group tend to shop predominantly during November and December. They spend considerably less time on administrative and informational pages, which may imply that they are more interested in specific products rather than exploring site-wide content. Additionally, visitors in this cluster are less likely to return compared to visitors in other clusters, with the exception of Cluster 3.

Cluster 5: This cluster appears to represent non-purchasers or those with low intent to buy. It has the highest exit rate and lowest revenue**.** Visitors in this cluster spend the least amount of time on administrative**,** informational, and product- related pages. Interestingly, despite their low purchase activity, 94.87% of visitors in this cluster are returning visitors, which may suggest they have been on the site previously but have not yet made a purchase.

Revenue Analysis: The chart shows that returning visitors (1) are more likely to make a purchase compared to non- returning visitors (0). The number of purchases among returning visitors is significantly higher, indicating a strong correlation between returning visitors and purchase likelihood. The revenue analysis focuses on understanding how different clusters contribute to overall sales and visitor behaviors. After applying PAM (Partitioning Around Medoids) and k-means clustering on the scaled data, we assessed the purchase ratio (revenue per cluster) and visitor behaviors. The purchase ratio is computed as the total revenue from a cluster divided by the number of visitors in that cluster. This helps understand which clusters are contributing the most in terms of sales relative to their size.



Key Findings

- **Cluster 3** shows the **highest purchase ratio**, which is consistent with its profile of returning visitors who make quick purchase decisions. This suggests that returning customers in Cluster 3 are the most valuable in terms of revenue.
- **Cluster 4**, while not the largest, also has a high purchase ratio, indicating that visitors in this group, especially during the holiday season (November and December), contribute significantly to overall revenue. However, their lower likelihood of returning suggests that their purchasing behavior is more seasonal.
- **Cluster 5**, with its low purchase ratio and high exit rate, represents visitors who likely do not intend to purchase, and their contribution to revenue is minimal.

We performed a Kruskal-Wallis test to statistically assess whether there is a significant difference in the purchase ratio across the clusters. This non- parametric test helps to determine if the differences in the revenue contribution of the clusters are significant.

Kruskal-Wallis Test

- **Null Hypothesis (H_0)**: There is no significant difference in the purchase ratio across the clusters.
- **Alternative Hypothesis (H_1)**: There is a significant difference in the purchase ratio across the clusters.

If the test returns a **p-value** less than **0.05**, we reject the null hypothesis and conclude that there are significant differences in the purchase ratios of the clusters. Based on the analysis, it was found that clusters with higher purchase ratios, such as Cluster 3, are significantly different from clusters like Cluster 5, which show low purchase behavior.

Descriptive Analysis: In this analysis, we explored the distribution and characteristics of the clusters to understand visitor behaviors more comprehensively. Here's a breakdown:

- **Cluster 0** has the highest number of returning visitors, indicating high loyalty and repeat engagement.
- **Cluster 3** stands out due to its high purchase ratio, signifying that visitors in this group tend to make purchases more frequently. These visitors spend minimal time on administrative and informational pages, highlighting a strong intent to buy once they arrive on the site.
- **Cluster 5** represents visitors who appear to be non-purchasers, as shown by their low revenue and high exit rates. Despite being returning visitors, they are less engaged with the product-related content on the site, which may suggest they are less likely to make a purchase in the future unless their needs are better addressed.

CONCLUSION

In this project, we used K-Means and PAM (Partitioning Around Medoids) algorithms for customer segmentation based on visitor behavior and revenue patterns. K-Means was chosen initially due to its simplicity and efficiency, providing quick insights into distinct visitor clusters. It helped identify six clusters, revealing key trends such as the high purchase ratio in Cluster 3 and low revenue in Cluster 5. Afterward, PAM was used to refine the clusters and validate the results, offering more robustness against outliers. Through revenue analysis, we found significant differences in the purchase behavior across clusters, with certain segments contributing more to overall revenue. This analysis, supported by the Kruskal- Wallis test, highlighted how visitor engagement impacts revenue generation. In conclusion, K-Means proved to be the more suitable algorithm for this dataset due to its scalability, while PAM was valuable for fine- tuning the results. The findings from this analysis can inform targeted marketing strategies and improve customer engagement. Future work could involve exploring other clustering algorithms, incorporating more features, and applying predictive models for further revenue optimization.

FUTURE WORK

Future work for this project includes improving the clustering algorithms by experimenting with different methods like DBSCAN or hierarchical clustering for better accuracy. Additionally, we can explore incorporating more features such as customer demographics or purchase history to enhance the segmentation process. Real-time data processing and prediction models could be added to make the system more adaptive and responsive to customer behavior changes. Finally,

optimizing the model for scalability can ensure it works effectively with larger datasets.

REFERENCES

Zavali, M. E. Lacka, and J. de Smedt, "Shopping Hard or Hardly Shopping: Revealing Consumer Segments Using Clickstream Data," IEEE Xplore, Apr. 2020.

Fangchun, Y. W. Shangguang, L. Jinglin, L. Zhihan, and S. Qibo, "Customer Segmentation of E-commerce Data Using K-means Clustering Algorithm," IEEE Xplore, Oct. 2020.

Kumar, R. S. M. K. Kumari, and S. R. Sharma, "Evaluation of Clustering Algorithms for Customer Segmentation," SpringerLink, Nov. 2021.

Suresh V. V. and P. K. Ranjan, "Performance Comparison of Clustering Algorithms for E- commerce Customer Segmentation," MDPI, Jan. 2022.

Gupta P. R. and T. S. Sharma, "Comparative Analysis of K-means, DBSCAN, and PAM for Customer Segmentation," SpringerLink, May 2022.

Kumar, P. G. M. S. Jain, and A. N. Gupta, "Application of K-means Clustering in Data Mining," ScienceDirect, Feb. 2018.

Kumar, N. S. R. R. Sharma, and K. Y. Bansal, "A Review on Customer Segmentation Techniques in Marketing," MDPI, Dec. 2020.

Singh, M. P. S. K. Thakur, and A. Kumar, "Customer Behavior Analysis Using K-means Clustering," IEEE Xplore, Jun. 2021.

Gpta S. R.and A. D. Singh, "A Comprehensive Survey of Clustering Algorithms," SpringerLink, Jan. 2022.

****