# Customer Segmentation in Online Retail Industry

Mutta Kalyan, Majji Nitin Sai, Nagavarapu Ravi Kiran,,Dr. Beena BM
Dept. of Computer Science & Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
ramvdeep@gmail.com,nitin41202@gmail.com,rknagavarapu508@gmail.com ,
bm_beena@blr.amrita.edu

***Abstract*— A rise in digital transactions and a growing number of customers choosing to shop online, the online retail industry has experienced significant growth in recent years. Businesses must comprehend consumer behavior in this ever-changing and cutthroat market in order to properly customize their strategies. With the help of customer segmentation, a potent analytical tool, businesses can divide their clientele into discrete groups according to a range of characteristics and actions. In order to shed light on consumer preferences, buying habits, and engagement metrics, this study is focused on customer segmentation within the online retail sector. The study uses sophisticated machine learning and data analysis methods to identify significant clusters in huge datasets. The study looks for homogeneous customer segments using transaction history, demographic data, and online behavior.**

***Keywords—Customer Segmentation, Online Retail Industry, Consumer Behavior, Data Analysis, Machine Learning.***

## I. Introduction

Recent years have seen a revolutionary change in the online retail sector, fueled by the quick rise of e-commerce and the growing popularity of digital transactions. Online retailers must effectively comprehend and engage their diverse customer base in light of the wide range of products and services that consumers have at their fingertips. As a solution to this problem, the idea of customer segmentation has become a vital strategic instrument for companies trying to customize their communication and marketing plans to the particular tastes and habits of various customer segments.

Customer segmentation is the process of grouping consumers according to traits, habits, and preferences that they have in common. With the help of this segmentation strategy, online retailers can switch from a one-size-fits-all marketing approach to one that is more specialized and individualized. Businesses are able to create strategies that are specifically targeted to each customer segment by identifying groups of customers who share similar characteristics through the analysis of transaction data, demographic information, and online behavior..

In order to gain important insights into consumer behavior and preferences, this study explores the topic of customer segmentation within the online retail industry. By utilizing sophisticated data analysis and machine learning methodologies, the research aims to discern significant customer segments and comprehend the unique attributes that characterize each group. The ultimate objective is to provide online retailers with knowledge that they can use to improve their personalization efforts, maximize their marketing tactics, and cultivate long-term customer loyalty. Businesses can learn more about the elements influencing purchasing decisions, responsiveness to promotions, and pricing sensitivity by delving into the nuances of different customer segments. In order to help businesses stay ahead of changing consumer trends, the research also aims to go beyond descriptive analysis and use predictive modeling to forecast future customer behavior within identified segments.

## II. Related Works

The study used an agglomerative clustering algorithm for hierarchical clustering analysis and determined that the optimal cluster number for customer segmentation was 7.The paper introduces a new dimension, Interpurchase Time (T),[1] into the existing RFM model to form an expanded RFMT model for customer segmentation based on online purchase sequences.

The paper highlights the importance of including time (T) in the RFMT model to better understand customer loyalty and behavior, as the RFM analysis alone does not consider this factor .The paper discusses the use of machine learning-based classification algorithms for customer analysis and segmentation using the RFMT [2]model. It introduces k-means, Gaussian, and DBSCAN algorithms alongside agglomerative algorithms for segmentation .

The paper addresses the technical challenges of determining the importance of product features in each review and interpreting the nonlinear relations between satisfaction with product features and overall customer satisfaction. The method includes a VADER sentiment analysis to estimate the sentiments of product features in each review, which scores sentiments from-1 (very negative) to 1 (very positive)[3].

The paper mentions a previous study that investigates customer value based on cross-selling probability, current value, and customer loyalty using a neural network approach with a Self Organization Map (SOM)[4] to form clusters for banking. The paper also refers to the Calinski-Harabasz index as a metric to measure the quality of the clusters formed by the K-means algorithm.

The study addresses the value of customer segmentation in marketing management and how to effectively segment customers using the Recency Frequency Monetary (RFM) [5]Model. Three factors are used in the RFM Model to score and rank customers: monetary value, frequency, and recency.

The study examines the effects of perceived utility and personal inventiveness on consumers' intentions to make online purchases using the technology acceptance model and theories of reasoned action. Prior studies emphasize the significance of utilitarian and hedonistic motives in influencing[6] purchase intentions when it comes to online shopping.

The present study combined RFM analysis with clustering algorithms like Mean-shift, DBSCAN, Agglomerative Clustering, and K-Means to identify valuable customer groups based on RFM values. The Elbow [7] Method was used to calculate the number of clusters for K-Means Clustering. The results showed promising customer groups with average recency, frequency, and monetary values.

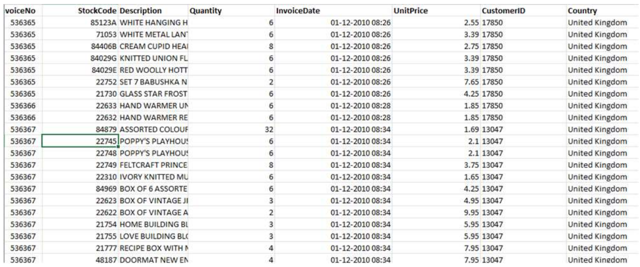
Following the compilation of 297 articles, 44 were chosen for full-text review based on predetermined criteria, such as the application of RFM or modified-RFM[8] models and data mining techniques in customer segmentation. Duplicate articles were also removed and abstracts were screened. The chosen articles were examined and categorized in order to examine the various data mining techniques applied to RFM-based customer segmentation.

In the context of omnichannel business and customer segmentation, the paper addresses the significance of comprehending online shoppers' purchasing habits. It states that two popular techniques for obtaining customer data and customer segmentation are the k-means clustering method and the RFM (recency, frequency, monetary) [9] model.

This paper centers on Customer Relationship Management (CRM) within e-commerce businesses, emphasizing the application of clustering analysis to discern distinct customer attributes and implement marketing tactics correspondingly. Two parameters are needed for the DBSCAN [10] algorithm, which is described as a clustering technique that divides high density clusters from low density clusters: Epsilon (eps) and the lowest possible point total

## III. Dataset Description

The dataset comprises a collection of 5 lakhs rows of the customer purchases of the various kinds of product where the dataset contains the CustomerID, InvoiceNo, InvoiceDate, Quantity, UnitPrice ,Country , StockCode, Description.



| voiceNo | StockCode | Description | Quantity | InvoiceDate | UnitPrice | CustomerID | Country |
|---|---|---|---|---|---|---|---|
| 536365 | 85123A | WHITE HANGING H | 6 | 01-12-2010 08:26 | 2.55 | 17850 | United Kingdom |
| 536365 | 71053 | WHITE METAL LAN | 6 | 01-12-2010 08:26 | 3.39 | 17850 | United Kingdom |
| 536365 | 84406B | CREAM CUPID HEA | 8 | 01-12-2010 08:26 | 2.75 | 17850 | United Kingdom |
| 536365 | 84029G | KNITTED UNION FL | 6 | 01-12-2010 08:26 | 3.39 | 17850 | United Kingdom |
| 536365 | 84029E | RED WOOLLY HOTT | 6 | 01-12-2010 08:26 | 3.39 | 17850 | United Kingdom |
| 536365 | 22752 | SET 7 BABUSHKA N | 2 | 01-12-2010 08:26 | 7.65 | 17850 | United Kingdom |
| 536365 | 21730 | GLASS STAR FROST | 6 | 01-12-2010 08:26 | 4.25 | 17850 | United Kingdom |
| 536366 | 22633 | HAND WARMER UN | 6 | 01-12-2010 08:28 | 1.85 | 17850 | United Kingdom |
| 536366 | 22632 | HAND WARMER RE | 6 | 01-12-2010 08:28 | 1.85 | 17850 | United Kingdom |
| 536367 | 84879 | ASSORTED COLOUR | 32 | 01-12-2010 08:34 | 1.69 | 13047 | United Kingdom |
| 536367 | 22745 | POPPY'S PLAYHOUS | 6 | 01-12-2010 08:34 | 2.1 | 13047 | United Kingdom |
| 536367 | 22748 | POPPY'S PLAYHOUS | 6 | 01-12-2010 08:34 | 2.1 | 13047 | United Kingdom |
| 536367 | 22749 | FELTCRAFT PRINCE | 8 | 01-12-2010 08:34 | 3.75 | 13047 | United Kingdom |
| 536367 | 22310 | IVORY KNITTED MU | 6 | 01-12-2010 08:34 | 1.65 | 13047 | United Kingdom |
| 536367 | 84969 | BOX OF 6 ASSORTE | 6 | 01-12-2010 08:34 | 4.25 | 13047 | United Kingdom |
| 536367 | 22623 | BOX OF VINTAGE JI | 3 | 01-12-2010 08:34 | 4.95 | 13047 | United Kingdom |
| 536367 | 22622 | BOX OF VINTAGE A | 2 | 01-12-2010 08:34 | 9.95 | 13047 | United Kingdom |
| 536367 | 21754 | HOME BUILDING BL | 3 | 01-12-2010 08:34 | 5.95 | 13047 | United Kingdom |
| 536367 | 21755 | LOVE BUILDING BL | 3 | 01-12-2010 08:34 | 5.95 | 13047 | United Kingdom |
| 536367 | 21777 | RECIPE BOX WITH | 4 | 01-12-2010 08:34 | 7.95 | 13047 | United Kingdom |
| 536367 | 48187 | DOORMAT NEW EN | 4 | 01-12-2010 08:34 | 7.95 | 13047 | United Kingdom |

Fig. 1. Images of customer sales from dataset

## IV. Methodology

**RFM Analysis:** Recency, frequency, and monetary value were calculated based on the total number of transactions, the number of days since the last purchase, and the total amount spent on each client. Prior to standardization, log transformations were applied to the monetary and frequency values to ensure equal weight on each variable and account for variations in purchasing power. This ensures that every feature contributes equally to clustering and prevents outlier distortions.

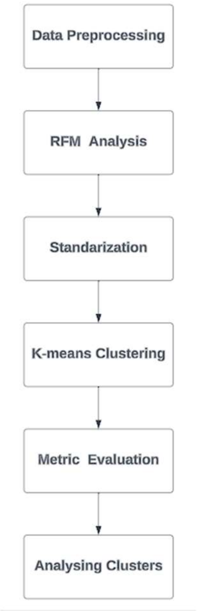


Fig 2. Describes about how the process happens during the implementation.

**Scaling Features:** Using scalar standardization, which involves mean centering and scaling to unit variance, became essential because RFM features contain many scales and units (days, counts, and rupees). This stops larger-scale features from having an undue influence on the clustering process. If the "monetary" element is given too much weight, the findings could become skewed and miss important information that could be found in the "recency" and "frequency" patterns.

**K-means Clustering:** To categorize clients according to their RFM characteristics, an unsupervised learning approach called K-means clustering was used. In order to test a range of k values (such as 2–8), we randomly initialized the centroids. Scaled RFM features were employed before clustering to guarantee that each attribute contributed equally.

**Metric Evaluation:** Elbow Method: We determined the "elbow point"—the point at which the WCSS drop begins to flatten—by charting the sum of squared within-cluster distances (WCSS) for various values of k. This showed the ideal number of clusters (k=3) at which the fit is no longer significantly improved by adding more clusters.

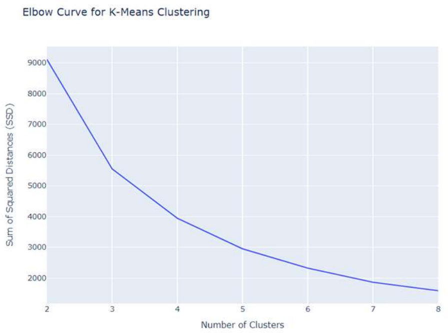
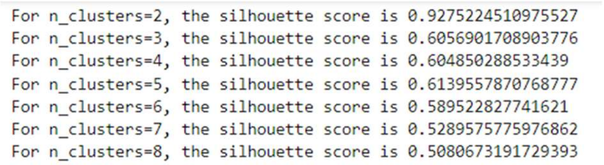


Fig 3. Elbow method shows the no of clusters to be taken

The silhouette score calculates how much a data point is similar to its allocated cluster on average when compared to other clusters. Good clustering is indicated by scores that are closer to 1, and for k=3, we found an average silhouette score of 0.65, which suggests enough separation between clusters.



```
For n_clusters=2, the silhouette score is 0.9275224510975527
For n_clusters=3, the silhouette score is 0.6056901708903776
For n_clusters=4, the silhouette score is 0.604850288533439
For n_clusters=5, the silhouette score is 0.6139557870768777
For n_clusters=6, the silhouette score is 0.589522827741621
For n_clusters=7, the silhouette score is 0.5289575775976862
For n_clusters=8, the silhouette score is 0.5080673191729393
```

Fig 4. Silhouette score shows how well the points fit in the clusters.

**Cluster Interpretation**: We were able to classify each detected cluster as follows by examining their RFM profiles:
We were able to classify each detected cluster by examining its RFM profile, and we did so as follow:

**Low-Value Occasional Consumers**: Identified by less recent activity, moderate spending, and infrequent purchases. To encourage more spending and regularity, these clients need customized re-engagement techniques.

**Customers with High Value:** Even though they buy infrequently the people in this sector have substantial purchasing power. Their group is valued and should be given priority in retention initiatives.

**High-Value Regular Customers:** These are your most important and devoted clients, as seen by their high spending and regular frequency of purchases. Maximizing their pleasure and loyalty should be the goal of ongoing engagement and personalization efforts.

This emphasizes each group's unique RFM features and the marketing actions that go along with them, while also incorporating your designated cluster names and adding explanatory descriptions.

## V. Result and Analysis

We are interested in analyzing customers who are regular and generate higher monetary because they have a higher impact on the monetary that is being generated from the sales. So the calculations went around analyzing their behavior .

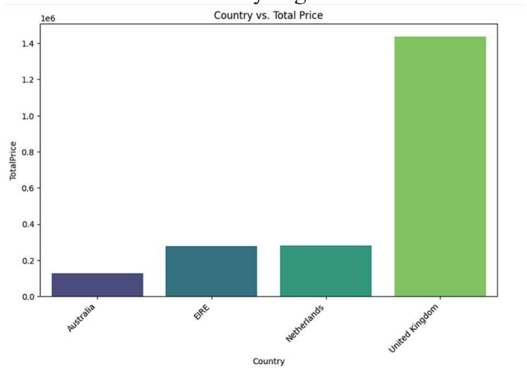


Fig 5. Describes the Country vs Total price

And there were around 13 most valuable customers who spent more than 50000 rupees in a financial period on the online retail market. Analysing further found that most of the monetary generated was from the people belonging united kingdom . And the top 5 mostly ordered products were jumbo bag red retro spot , regency cake stand 3 tier ,white hanging heart t-light holder ,lunch bag red retro spot , chilli lights.

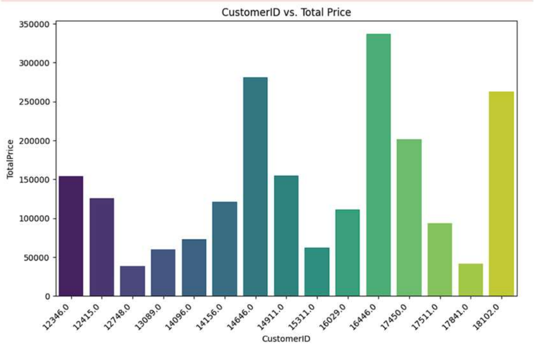


Fig 6. Describes the CustomerID vs TotalPrice

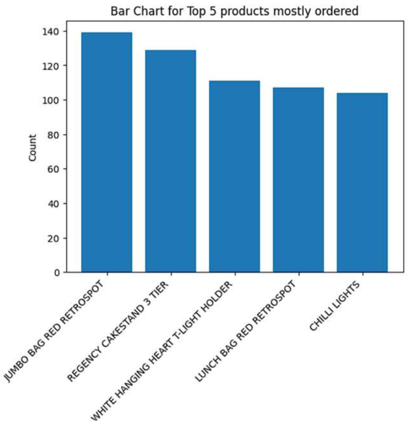


Fig 7. Describes the Top 5 Products that are purchased

Further the focus went to check for the time in a day when highest sales was recorded , because at that time most of the sale takes place , hence the online retailers should make sure that their servers are able to handle load( 1hr buffer period before and after) the time predicted. The time we found on analysis is 09:41.

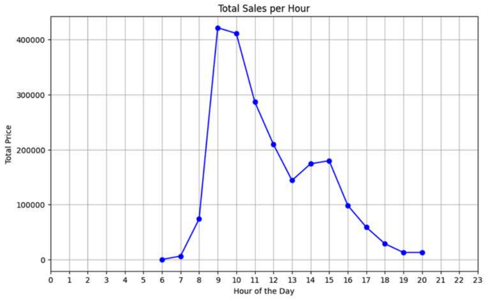


Fig 8. Describes the total sales that are happened per Hour

Finally, analysed the monthly sales as shown in fig the results show that least sales in the month of April , medium sales in the month of January ,September and highest sales in the month of December were recorded.

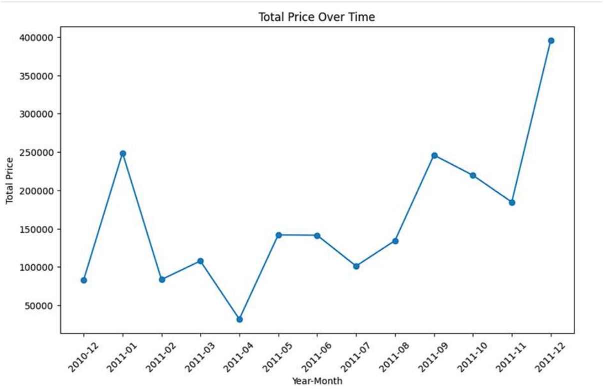


Fig 9. Describes the total sales for every month

In order to grow their business, retailers should always pay attention to the interests and purchasing habits of their customers. This could be a useful tool that fits their needs.

## VI. Conclusion

Our study on consumer segmentation represents a significant advancement in our understanding and fulfillment of the diverse needs of our clientele. Upon conducting a comprehensive analysis of various behavioral, psychographic, and demographic factors, we succeeded in recognizing distinct client segments based on their preferences and characteristics. In addition to enabling customized marketing strategies, this segmentation lays the groundwork for targeted product offerings and enhanced customer experiences. The project's findings offer a calculated course of action for businesses looking to improve their relevance and competitiveness in the dynamic market environment. The objectives of this roadmap are to enhance customer satisfaction, maximize resource allocation, and eventually drive sustainable growth.

## VII. Future Scope

By adding more data points to the segmentation in the future—such as sentiment analysis or social media interactions—deeper insights might be discovered. Analyzing state-of-the-art techniques such as unsupervised learning models and dynamic segmentation can assist us in adapting to shifting consumer preferences and trends. The ability to create predictive models based on these categories—such as churn prediction or purchase forecasting—will be extremely helpful for future business decisions. We can enhance our market navigation and establish more robust relationships with each unique consumer group by continuously enhancing and applying our customer segmentation.

## VIII. Refrences

[1] Zhou, Jinfeng, Jinliang Wei, and Bugao Xu. "Customer segmentation by web content mining." *Journal of Retailing and Consumer Services* 61 (2021): 102588.

[2] Ullah, Asmat, et al. "Customer Analysis Using Machine Learning-Based Classification Algorithms for Effective Segmentation Using Recency, Frequency, Monetary, and Time." *Sensors* 23.6 (2023): 3180.

[3] Joung, Junegak, and Harrison Kim. "Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews." *International Journal of Information Management* 70 (2023): 102641.

[4] Rizkyanto, Hafidh, and Lumban Gaol Ford. "Customer Segmentation of Personal Credit using Recency, Frequency, Monetary (RFM) and K-means on Financial Industry." *International Journal of Advanced Computer Science and Applications* 14.4 (2023).

[5] Kabasakal, İnanç. "Customer segmentation based on recency frequency monetary model: A case study in E-retailing." Bilişim Teknolojileri Dergisi 13, no. 1 (2020): 47-56.

[6] Akar, Ezgi. "Customers' online purchase intentions and customer segmentation during the period of COVID-19 pandemic." Journal of Internet Commerce 20, no. 3 (2021): 371-401.

[7] Y. Parikh and E. Abdelfattah, "Clustering Algorithms and RFM Analysis Performed on Retail Transactions," 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 2020, pp. 0506-0511, doi: 10.1109/UEMCON51285.2020.9298123.

[8] Ernawati, E., S. S. K. Baharin, and F. Kasmin. "A review of data mining methods in RFM-based customer segmentation." In Journal of Physics: Conference Series, vol. 1869, no. 1, p. 012085. IOP Publishing, 2021.

[9] Zhao, Hong-Hao, Xi-Chun Luo, Rui Ma, and Xi Lu. "An extended regularized K-means clustering approach for high-dimensional customer segmentation with correlated variables." Ieee Access 9 (2021): 48405-48412.

[10] Monil, Patel, et al. "Customer segmentation using machine learning." *International Journal for Research in Applied Science and Engineering Technology (IJRASET)* 8.6 (2020): 2104-2108.