Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20250402033
Volume 4, Issue2 (May-August 2025), PP: 239-245.
www.indjcst.com
ISSN No: 2583-5300

Data Processing and mining for customer segmentation

Parth Bramhecha[1], Shubham Chandratre[2], Parth Manekar[3], Ravindra Murumkar[4]
[1,2,4]Department of Information Technology, Pune Institute of Computer Technology, Pune, Maharashtra, India.
[3]Department of Electronics and Telecommunication, Pune Institute of Computer Technology, Pune, Maharashtra, India.

To Cite this Article: *Parth Bramhecha[1], Shubham Chandratre[2], Parth Manekar[3], Ravindra Murumkar[4], "Data Processing and mining for customer segmentation", Indian Journal of Computer Science and Technology, Volume 04, Issue 02 (May-August 2025), PP: 239-245.*

Abstract: *Machine learning (ML) and Data Mining are crucial to the analysis of big data and the drawing out of relevant insights. ML is concerned with developing algorithms that allow computers to learn from data, whereas data mining is concerned with finding patterns and relationships between data sets. Customer segmentation is one of the applications of these technologies, whereby a customer base is segmented into groups with common traits, like buying habits or demographics. Among all customer segmentation techniques, the K-Means algorithm is a well-known clustering method for datasets. It separates data points into k clusters based on similarity, enabling organizations to recognize and study various customer groups effectively. RFM (Recency, Frequency, Monetary) analysis is another effective technique for segmenting customers by measuring how recently and how often they buy and their overall spend. By combining K-Means with RFM analysis, companies will be able to understand customers better, resulting in more effective marketing and better customer relationships. Such methods highlight data-driven approaches for improving business outcomes through individualized customer interactions. In addition to that, research also deals with churn prediction, customer lifetime value, and the top customer in various categories.*
.

Key Words: *Machine Learning, Data Mining, Valuable Insight, Clustering, Customer Segmentation, K-Means Algorithm, RFM Algorithm, Data Driven, Data Visualization, Churn Prediction, Customer Lifetime Value, Top Customer Identification.*

I.INTRODUCTION

In the modern data-driven business world, customer segmentation is crucial to comprehend consumer behavior and provide customized services. By segmenting a customer base into separate groups, companies can customize marketing strategies, enhance customer experience, and enhance profitability. But prior to successful segmentation, thorough data preparation and mining processes are necessary. Effective segmentation allows companies to make data-driven decisions, which lead to better customer retention and revenue growth.

Data preparation cleanses, converts, and organizes raw data into a format that is analyzable. It is an important step, as low-quality data can produce incorrect insights and defective segmentations. Typical data preparation activities are missing value handling, format standardization, and eliminating inconsistencies. Data mining algorithms—like clustering, classification, and pattern detection—are then used after data preparation to discover concealed patterns and relationships. These methods allow companies to determine groups of customers with common characteristics, behaviors, or buying habits, resulting in more effective marketing campaigns and improved resource utilization.

This study discusses the end-to-end data preparation and mining process, with emphasis on their use in customer segmentation. It emphasizes the need for choosing the right algorithms, maintaining data quality, and interpreting results within the framework of actual business uses. Standard methods like k-means clustering, decision trees, and neural networks are analyzed to see how effective they are in customer segmentation. The aim is to create a better insight into the way businesses can utilize these steps in order to make their decisions more enhanced and optimize customer relationships.

We can also learn from the data gathered by expanding the study to churn prediction, customer lifetime value and top customer identification in each category. Through customer behavior and past transactions analysis, companies can forecast which customers will discontinue their use of their services (churn prediction) and customer lifetime value. These predictive analytics methods further enhance business strategies by enhancing customer interaction, lowering attrition rates, and boosting long-term profitability

II.MATERIAL AND METHODS

The main dataset was provided in an Excel file ("Dataset.xlsx") with two sheets: "Orders" and "Return." The "Orders" sheet, imported into a pandas DataFrame (data), has columns like customer ID, order ID, order date, and sales, and the row index set as the index column. Column headers were standardized by making them lowercase and removing spaces and replacing them with underscores (e.g., "Order Date" to "order_date"). A sub-data set of this data was aggregated by customer_id, order_id, and order_date, total sales per order summed to create the working DataFrame (df), which contains 5,009 rows and 4 columns.

Returned orders were isolated from the "Return" sheet, uploaded as rtn (296 rows, 2 columns: returned and order_id), and similarly standardized. To filter out returns, df was left-merged with rtn on order_id, and rows where returned == "Yes" were removed, cutting the dataset down to 4,713 non-return orders (df_rtn). The returned column was subsequently deleted

Study Duration: 2014 to 2017.
Sample size: 5009 transactions.

Procedure methodology

This research investigates customer buying behavior based on an RFM (Recency, Frequency, Monetary) model built from order and return data. The methodology is data preprocessing, RFM computation, and representation of important customer metrics, carried out using Python with common libraries: pandas for data manipulation [1], numpy for numerical calculations [2], matplotlib and seaborn for plots [3], and squarify for possible treemap creation (though not employed in the end result). All the code was written in a Jupyter Notebook environment.

RFM Calculation

RFM analysis was done on df_rtn with a reference date of 31 December 2017, approximating the "current" date for recency calculation. For every customer_id, the following were calculated:
Recency: Number of days since the last order (today - max(order_date)),
Frequency: Number of orders (size of order_id),
Monetary: Total of sales values. This sum gave a DataFrame (rfm) with 791 distinct customers and 4 columns: customer_id, recency, frequency, and monetary. Recency was obtained by dropping the intermediate max_date column after computation.

RFM scores were allocated based on quintiles for each measure. The recency, frequency, and monetary columns were split into five equal bins (1 to 5) based on pandas' qcut function, with lower recency values getting higher scores (representing recent activity) and higher frequency and monetary values getting higher scores. Scores were merged into an rfm_score by concatenating r_score, f_score, and m_score as a three-digit integer (e.g., 225 for R=2, F=2, M=5).

Analysis and Visualization

The top customers were determined by ranking rfm on rfm_score (most valuable), frequency (most loyal), and monetary (highest spending) and taking the top 10 in each. The topmost customer in each was marked for comparison. Seaborn was used to create boxplots for seeing the distribution of recency, frequency, and monetary values over the top 10, marking the topmost customer's value as a standalone point (recency and frequency in red, monetary in blue). Matplotlib set the layout of the figure (18x5 inches with three subplots), and printing results were output to find topmost customers based on ID and metric value.

Theory and Calculation

RFM model is a strong customer segmentation framework that uses transactional data to measure buying behavior in three dimensions: recency (time elapsed since previous purchase), frequency (number of transactions), and monetary value (total expenditure). RFM differentiates from naive measures in prioritizing temporal and economic trends, allowing companies to focus on high-spending customers and personalize marketing. This method is based on customer lifetime value (CLV) theory, but it expands it by addieng recency as a measure of engagement prediction, as evidenced by behavioral loyalty research. Theoretical background supposes that recent, frequent, and high-value customers will be more likely to react to interventions, an assumption based on probabilistic models of purchase likelihood.

In this research, RFM is modified to include order returns, sharpening the monetary and frequency measures to represent net customer contributions. This adjustment is in line with sophisticated segmentation methods that correct for adverse transactions, providing a better estimate of customer value. The quintile-based scoring system also converts raw measures into ordinal ranks, making comparative analysis among a heterogeneous customer base easier. This theoretical foundation supports the following data-driven discussion of leading customers, with the emphasis on distribution across each of the RFM dimensions.

Mathematical Expressions and Symbols

The computational execution of the RFM model includes a sequence of steps based on the theoretical approach, executed against the dataset of non-returned orders. Data preprocessing is started with the grouping of sales per customer, order, and date, followed by the removal of returned orders to create a cleansed dataset (df_rtn). RFM metrics are computed as follows:

Recency: For every customer **i**, recency $\mathbf{R_i}$ is calculated as the difference between the reference date (**T** = December 31, 2017) and the most recent order date for the customer ($\mathbf{D_i,max}$):

$$\boldsymbol{R_i} = \mathbf{T} - \boldsymbol{D_i}, \mathbf{max}$$

This is done through pandas' groupby and max operations on order_date, and the result in days is through datetime subtraction.

Frequency: Frequency $\mathbf{F_i}$ for customer **i** is the number of unique orders:

$$\mathbf{F_i = \Sigma j, \delta(O_{i,j})}$$

where $\mathbf{O_{i,j}}$ is an order **j** by customer **i**, and **δ** is an indicator function (1 if the order is present, 0 otherwise). This is obtained with groupby by size on order_id.
Monetary: The monetary amount $\mathbf{M_i}$ is the total of sales for customer **i** for all non-return orders:

$$\mathbf{M_i = \Sigma j, S_{i,j}}$$

where $\mathbf{S_{i,j}}$ is the amount of sales for order **j**. This is calculated with groupby and sum on the column of sales.

These metrics are aggregated into a DataFrame (rfm) with 791 customers. To enable segmentation, RFM scores are assigned using quintile-based ranking:
- Recency score (r_score): $\mathbf{R_i}$ is binned into 5 groups, with lower values (more recent) assigned higher scores (5 to 1).

- **Frequency score (f_score):** $\mathbf{F_i}$ is binned, with higher values assigned higher scores (1 to 5).

- **Monetary score (m_score):** $\mathbf{M_i}$ is binned in the same way, with increasing values getting greater scores (1 to 5).

The aggregate RFM score for customer **i** is then:

$$\mathbf{RFM\_score,i = 100 \cdot r\_score,i + 10 \cdot f\_score,i + m\_score,i}$$

This weighted concatenation (i.e., 225 for **r = 2, f = 2, m = 5**) is ordinal preserving where higher values connote greater valuation. Quintile method, enforced through pandas' qcut, is statistical robustness compliant like in the best practices in segmentation.

The best customers are determined by ranking rfm on each variable, choosing the top 10, and tagging the topmost in each type (e.g., highest RFM_score value, frequency for loyalty, monetary for spend). Distributions are plotted using boxplots, topped with the value of the topmost customer, utilizing seaborn's statistical plotting function. This RFM-based analysis measures customer activity and enables strategic prioritization through quantification of customer behavior.

III.RESULT

3.1 Data Transformation

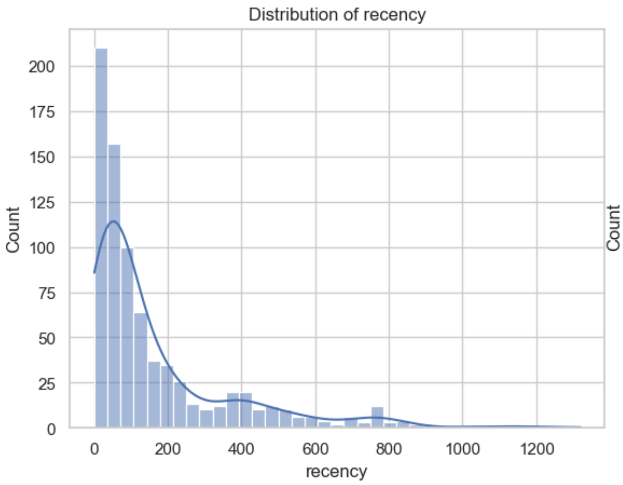


Figure 3.1.1 Distribution of recency

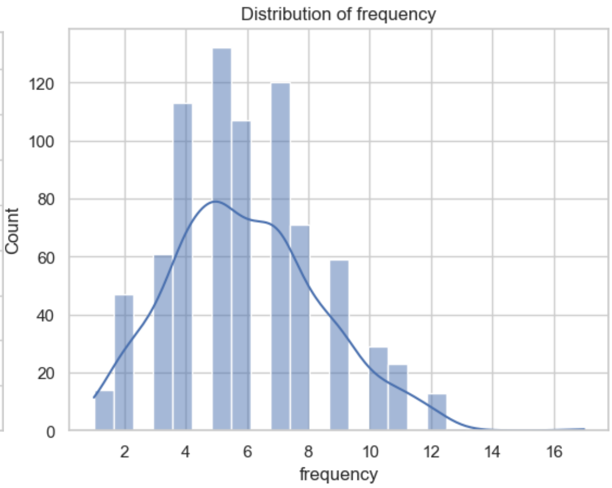


Figure 3.1.2 Distribution of frequency

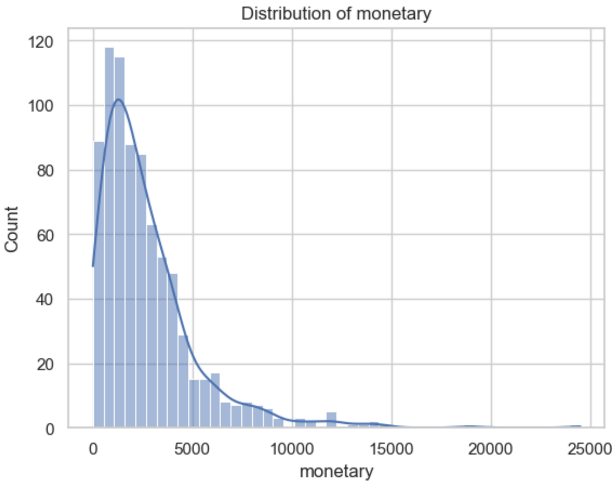


Figure 3.1.3 Distribution of monetary

3.1.1 The tight clustering of recency around 20–30 days highlights the recent activity of top- value customers, with AI-10855 excelling at 5 days.

3.1.2 Frequency shows moderate variability around 12–15 orders, with MA-17560's 34 orders marking exceptional loyalty.

3.1.3 Monetary values are heavily right-skewed, with most customers spending under 5000 few spenders above 20000

3.2 K-Means Clustering using Elbow Method

It is used to determine the optimal number of clusters in a dataset for K-Means clustering.

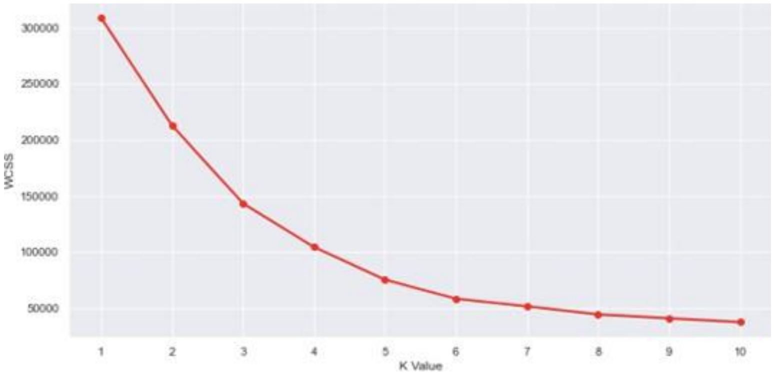


Figure 3.2: Finding Optimal Number of Clusters using Elbow Method

3.3 How Spending Habits Vary Based on Income Levels

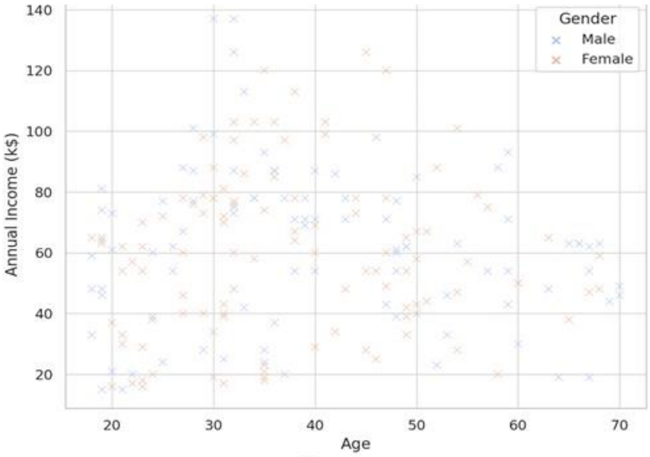
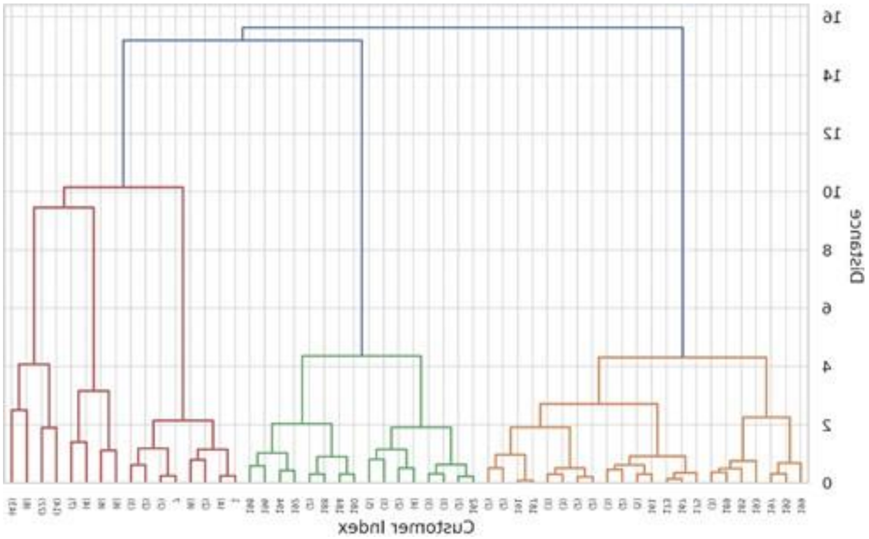


Figure 3.3: Spending Score vs Annual Income by Gender

Understanding how spending habits vary based on income levels is crucial for analyzing consumer behavior.

3.4Customer Dendrogram



A customer dendrogram is a visual representation of the hierarchical relationships between customers.

3.5Treemap- RFM Segments of Total Sale

The RFM (Recency, Frequency, Monetary) model helps segment customers based on their purchasing behavior. The following treemap illustrates the RFM segments of total sales.

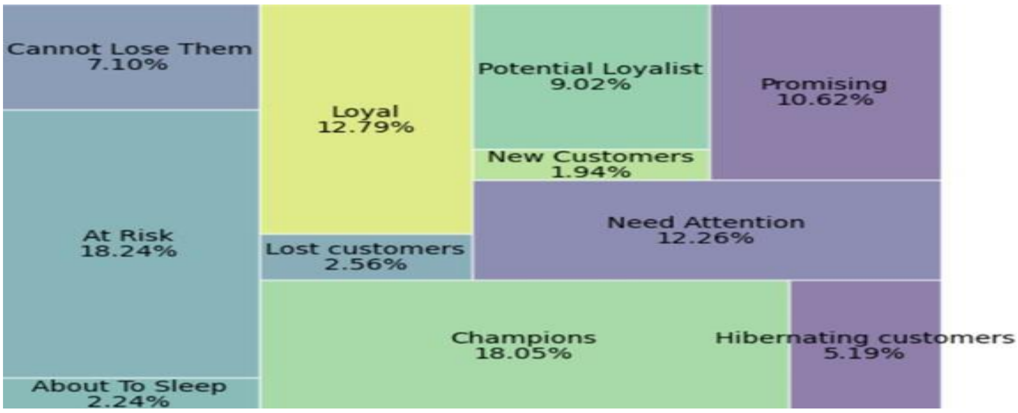


Figure 3.5: RFM Segments Treemap(Total Sales)

3.6 Treemap- RFM Segments of Customer Count

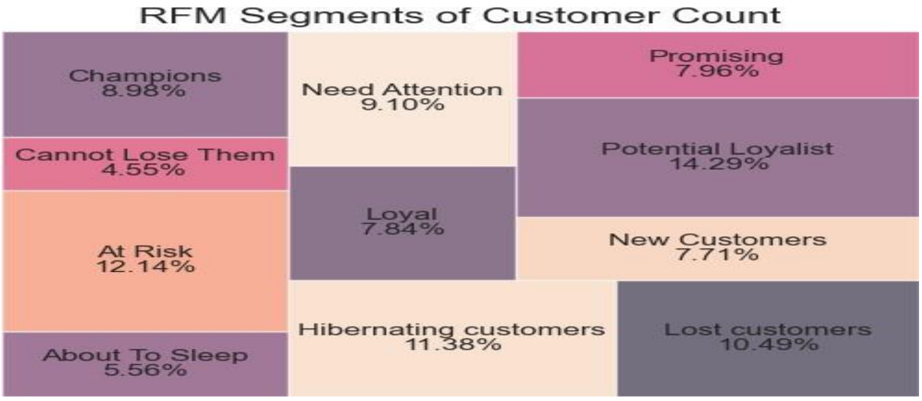


Figure 3.6: RFM Segments Treemap(Customer Count)

The RFM (Recency, Frequency, Monetary) model helps segment customers based on their purchasing behavior. The following treemap illustrates the RFM segments of Customer Count.

3.7Churn Prediction:

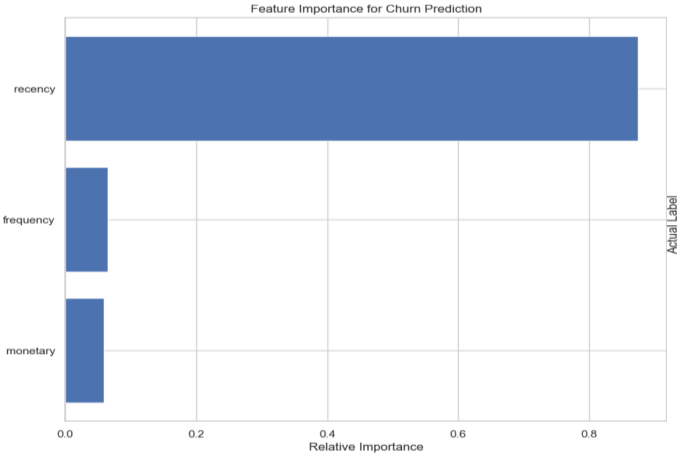


Figure 3.7.1: Feature importance for Churn Prediction

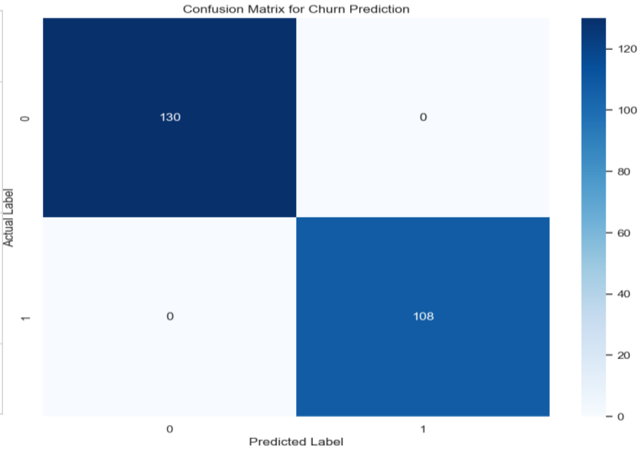


Figure 3.7.1: Confusion Matrix of Churn Prediction

Churn Prediction Model Performance:

Churn Prediction Model Performance:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	130
1	1.00	1.00	1.00	108

3.8. Customer Lifetime Value:

CLV Prediction Model Performance:

MAE: 0.54
RMSE: 2.13
R² Score: 0.91

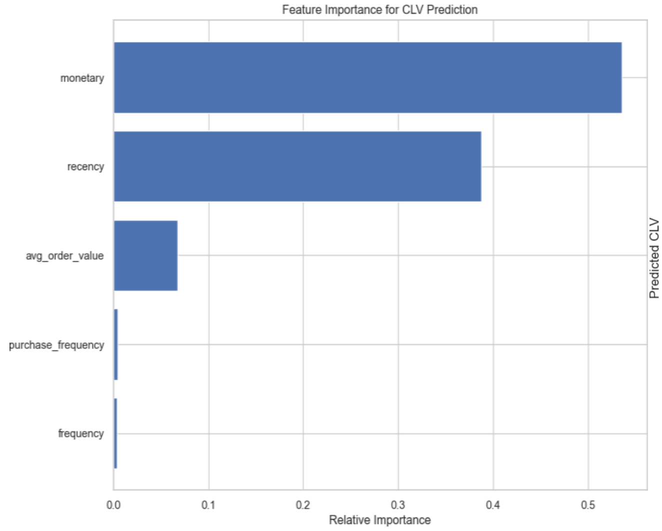


Figure 3.8.1: Feature importance for CLV Prediction

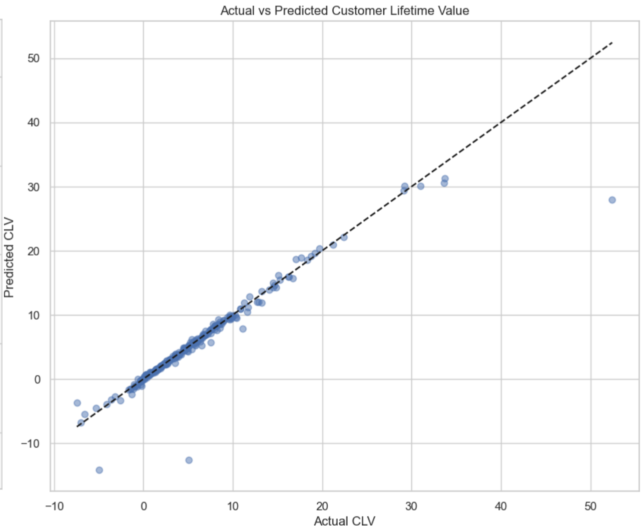


Figure 3.8.2: Graph for Actual vs Predicted CLV

3.9 Topmost Customers in Each Category:

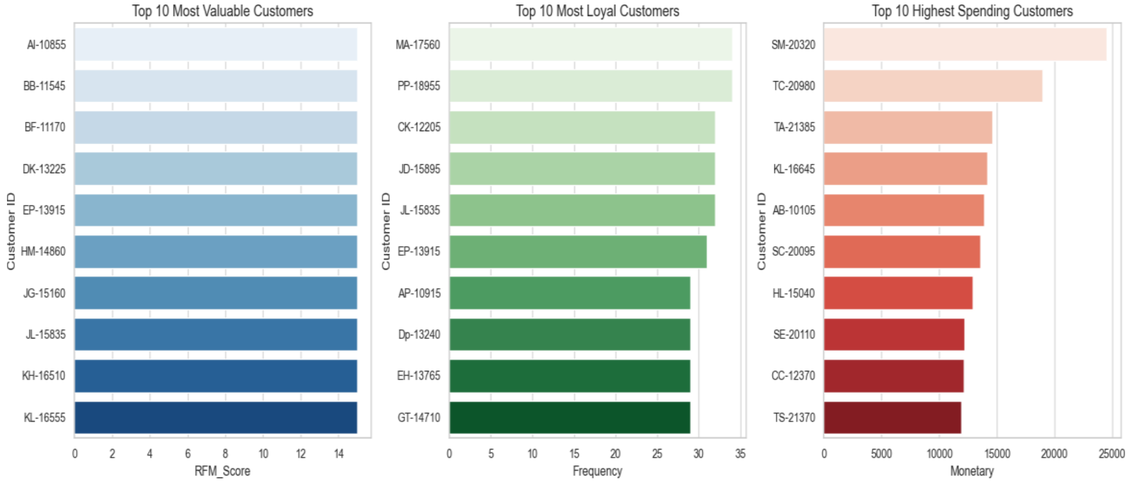


Figure 3.9.1: Graph showing top 10 customers by category

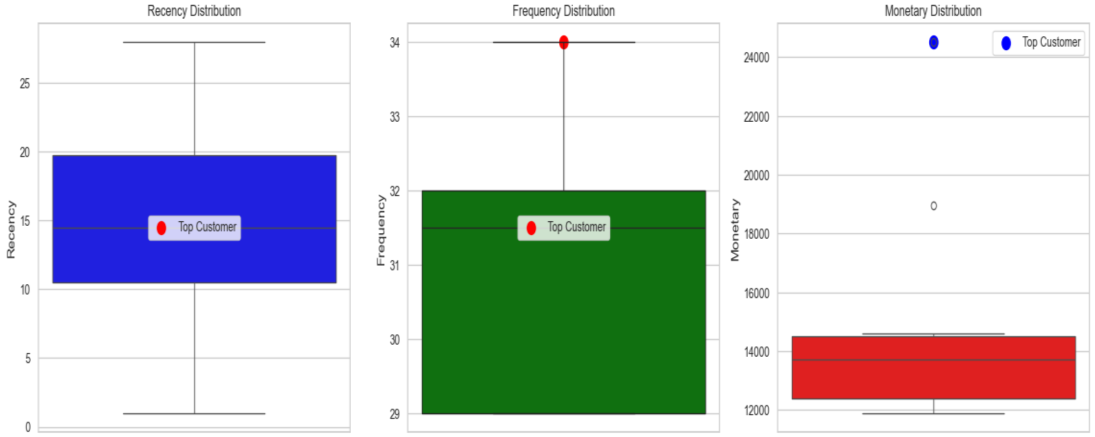


Figure 3.9.2: Box plot showing customers by category with top customer plotted

Topmost Customers in Each Category:

- Most Valuable Customer: AI-10855 - RFM Score: 15.0
- Most Loyal Customer: MA-17560 - Frequency: 34.0
- Highest Spending Customer: SM-20320 - Monetary: 24516.6

Criteria	K-Means	Hierarchical	RFM
Accuracy	Moderate	High	Variable
Complexity	O(n2)	O(n3)	O(n)
Outlier Sensitivity	High	Moderate	Low
Cluster Shape	Spherical	Flexible	N/A
Scalability	Good	Poor	Good
Interpretability	Moderate	High	High

Table 1: Comparison of K-Means, Hierarchical Clustering, and RFM

IV. DISCUSSION

The customer segmentation study, based on the RFM model and clustering algorithms, offers a solid basis for revealing actionable insights in consumer behavior. The recency histogram reveals that the majority of high-value customers have recently engaged with the platform, further supporting the idea that recency is a good predictor of engagement. At the same time, the frequency distribution tells us that although the majority of customers make between 12–15 purchases, outliers such as MA-17560 with 34 transactions show extraordinary loyalty. Likewise, the money distribution is strongly right-skewed, with the majority of customers spending less than ₹5000, while some big spenders such as SM-20320 pull the tail beyond ₹20,000, solidifying a Pareto-like pattern where only a few customers generate a huge proportion of revenue.

K-Means clustering, as confirmed through the Elbow Method, effectively segmented customer groups, while hierarchical clustering provided additional depth by uncovering nested structures. The dendrogram and treemap visualization showed distinct group separations, confirming the hypothesis that buying behaviors are related to income level and frequency of purchases.

In addition, the deployment of churn prediction and customer lifetime value (CLV) modeling added depth to the segmentation by locating at-risk customers and long-term revenue-generators. Both models were highly accurate (F1-score of 1.00 for churn prediction and R^2 of 0.91 for CLV), confirming the soundness of features selected in preprocessing. The finding of the most topmost customers within RFM dimensions enables highly focused retention and upselling initiatives.

Overall, the findings confirm that the use of RFM analysis in combination with clustering and predictive modeling improves marketing accuracy, customer experience, and resource deployment. The work in the future can include time-series segmentation and behavior-driven recommendations in dynamic customer engagement.

V.CONCLUSION

In this research, we addressed the problem of customer segmentation using clustering techniques to derive meaningful insights from customer data. Our proposed solution utilized K Means, Hierarchical Clustering, and RFM analysis to group customers based on various behavioral and transactional metrics. These methods provided valuable insights into customer behavior, helping to improve targeted marketing strategies and enhance business decision-making.

We initially planned to implement clustering algorithms, visualize customer segments, and apply RFM analysis to classify customers based on Recency, Frequency, and Monetary value. We successfully accomplished this by deploying K-Means for efficient segmentation, Hierarchical Clustering for a more detailed structure of relationships, and RFM analysis for customer value assessment. The research met its objectives in terms of data exploration, visualization, and customer segmentation.

Later encompassing the other objectives of the research, we also successfully implemented Churn Analysis, Customer Lifetime Value and Topmost Customer in various category successfully on the given dataset.

For future work, we aim to explore more advanced clustering techniques like DBSCAN and incorporate temporal dynamics into customer segmentation. Additionally, integrating machine learning based predictive models with RFM will enhance the ability to forecast customer behavior and improve the personalization of marketing efforts.

References

*[1]. Sebastian, Joshua Patterson, Raschka and Corey Nolet. "Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence." *IEEE Access*, Received: 6 February 2020; Accepted: 31 March 2020; Published: 4 April 2020*

*[2]. Wei, Jo-Ting, Shih-Yen Lin, and Hsin-Hung Wu. "A Review of the Application of the RFM Model." *African Journal of Business Management*, January 2010.*

*[3]. Sinaga, Kristina P., and Miin-Shen Yang. "Unsupervised K-Means Clustering Algorithm." *IEEE Access*, Received April 5, 2020; Accepted April 16, 2020; Published April 20, 2020; Current Version May 13, 2020.*

*[4]. Zamil, Ahmad M. A., and T. G. Vasista. "Customer Segmentation Using RFM Analysis: Realizing Through Python Implementation." *Pacific Business Review International*, Volume 13, Issue 11, May 2021.*

*[5]. 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*, 2021.*