Customer Segmentation Using Clustering Algorithms on Online Retail Data

Dilhara Disanayaka
Department of Computer Science & Engineering
University of Moratuwa, Sri Lanka
Email: dilhara.22@cse.mrt.ac.lk

***Abstract*—This study explores customer segmentation in online shopping using clustering algorithms to be applied on purchase behavior data. We experimented with three algorithms—K-means, hierarchical clustering, and DBSCAN—on RFM (Recency, Frequency, Monetary) features as input data. Hierarchical clustering with single linkage gave the best of the methods tried with a silhouette score of 0.945, while for K-means it was 0.581 and that for DBSCAN was 0.470. The research produced three sets of customers with varying shopping behaviors. These are findings that can be used to guide the formulation of customer relationship and focused marketing strategies.**

I. Introduction

Retailers in recent times gather huge amounts of data from customer purchases. This information provides insights into how consumers shop and prefer things. With segmentation of customers exhibiting similar behavior, companies can develop more successful marketing campaigns.

We use RFM (Recency, Frequency, Monetary) analysis and three clustering methods to segment customers in this study. We then compare results to find out which method performs better and analyze the business insights derived from the segments.

II. Methodology

A. Feature Engineering

We used the Online Retail Dataset from UCI Repository. The dataset contains transactions for a UK-based online retailer. The original dataset had 541,909 records with 8 columns.

we used several data pre-processing techniques. First, removed records with missing CustomerID values. This reduced the dataset to 406,829 records—next filtered data rows for United Kingdom customers only. All the cancelled orders were filtered and removed, leaving 354,345 records.

RFM features were calculated:

- **Recency**: Days since last purchase from the maximum date (2011-12-09)
- **Frequency**: Total number of distinct purchases per customer
- **Monetary**: Total purchase value (Quantity × UnitPrice)

The RFM analysis created 3,920 customer records. To remove the outliers used the IQR method and the final dataset contained 3,538 customers. Features were standardized using StandardScaler to ensure equal weighting.

The RFM feature distributions are shown in Figure 2, which reveals the characteristics of our customer data. The 3-D scatter plot analysis in Figure 1 illustrates the relationships between different RFM components. The correlation matrix shows (Figure 3), there are negative correlations between Recency and other features, while between Frequency and Monetary has a positive correlation (0.509).

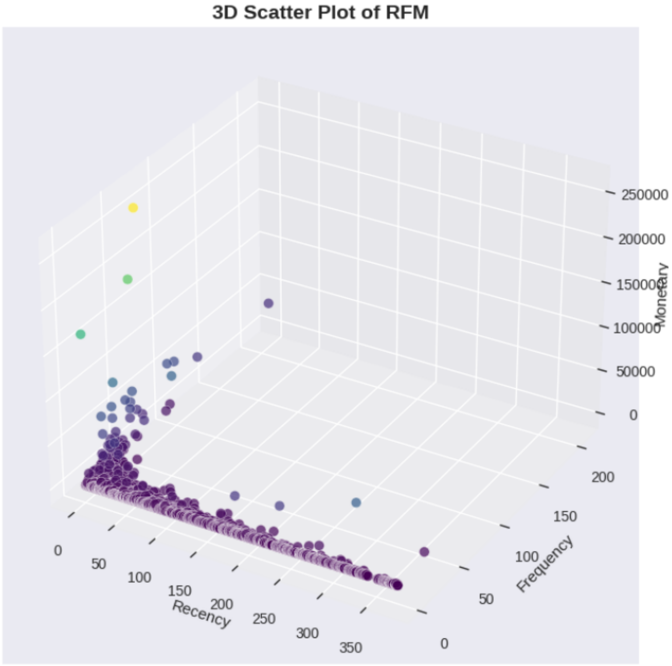

Fig. 1. 3D Scatter plot of RFM

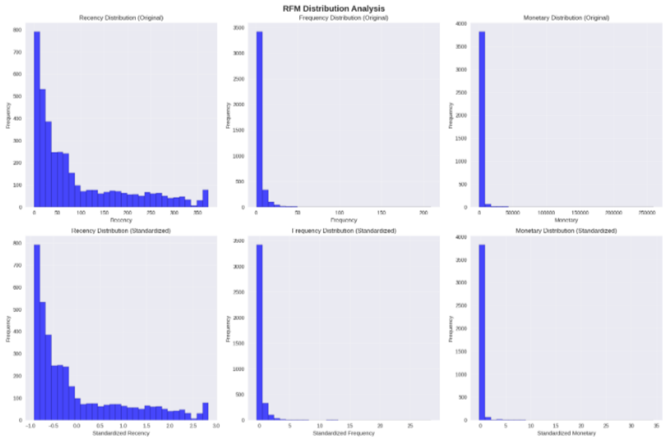

Fig. 2. Distribution of RFM Features

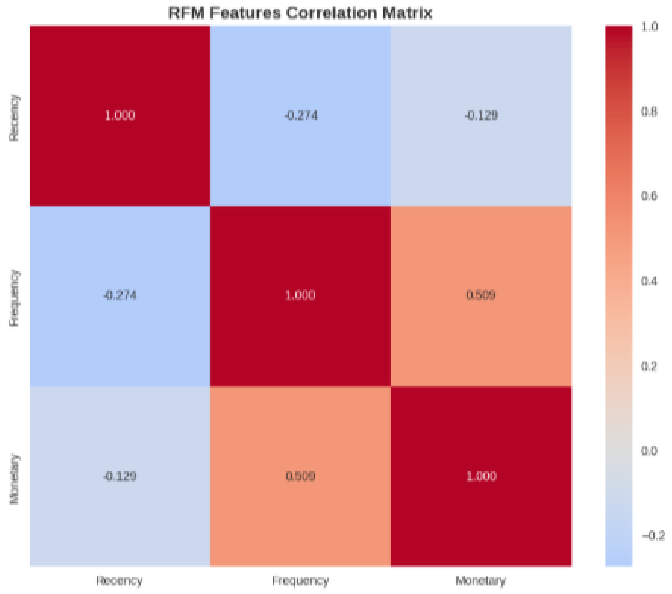

Fig. 3. Correlation Matrix of RFM Features

B. Clustering Methods

1) K-means Clustering: To find the optimal number of clusters we used the elbow method. The elbow curve and silhouette scores are presented in Figure 4. Based on these plots, we selected k=3 as the optimal cluster count because as this value provides a good balance between model complexity and clustering quality.

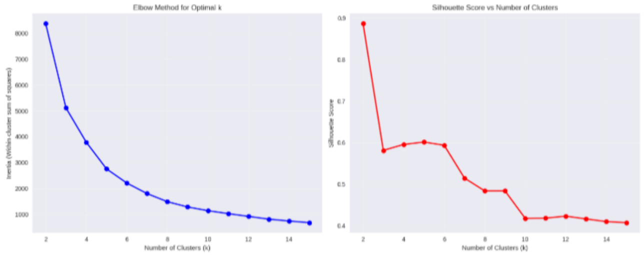

Fig. 4. Elbow Method for K Selection

2) Hierarchical Clustering: We applied three hierarchical clustering linkage methods: single, complete, and average. The corresponding dendrograms are presented in Figure 5. Among these, the single linkage method achieved the highest silhouette score, indicating stronger and more distinct cluster separation within our dataset.

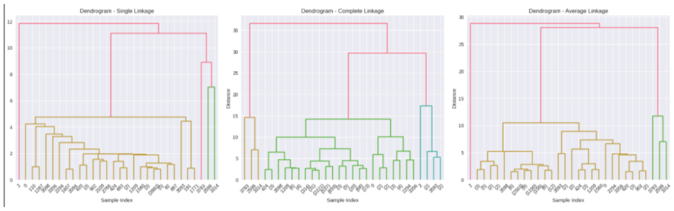

Fig. 5. Dendrograms with Single, Complete and Average Linkage

3) DBSCAN: To select suitable epsilon values, we considered the k-distance plots (Figure 6). The analysis indicated a potential epsilon range between 0.074 and 0.167. By trial and error of various epsilon and min$_samplesvalues, theoptimalclusteringperformancewasobtainedw$ $0.181andmin_samples = 10$.

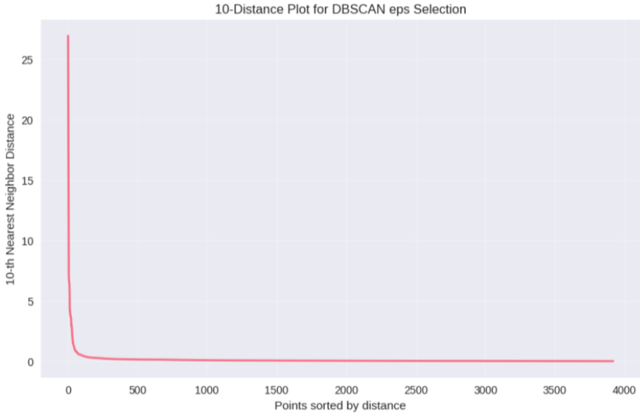

Fig. 6. K-Distance Plot for DBSCAN Parameter Selection

C. Visualization

We applied Principal Component Analysis (PCA) to map the clustering results to a two-dimensional space for visualization. Figure 7 illustrates how each of the clustering algorithms divided the customer data. The first two principal components captured a large percentage of the total variance, so the 2D representation is a good reflection of the main structure in the data set.

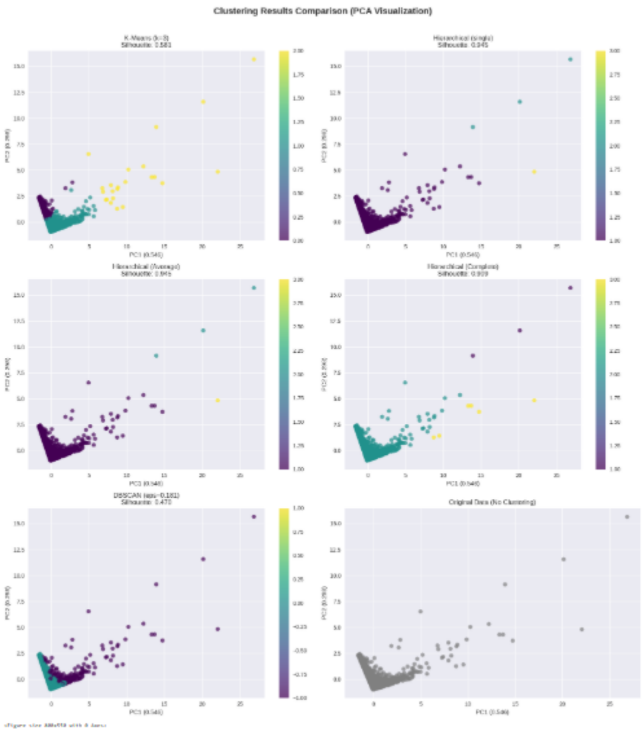

Fig. 7. Clustering Results Comparison using PCA

1) Subplot Interpretations: Each subplot in Figure 7 reveals distinct clustering characteristics:

K-Means (k=3, Silhouette = 0.581):K-Means produced three clusters with clear separation but some overlap at the boundaries. The moderate silhouette score reflects reasonable clustering quality, though results may be less robust compared to hierarchical methods.

Hierarchical – Single Linkage (Silhouette = 0.945: This method achieved the highest silhouette score, indicating very strong separation. However, the dendrogram suggests "chained" structures, with most points falling into one large cluster. While statistically strong, this may reduce practical usefulness in business segmentation.

Hierarchical – Average Linkage (Silhouette = 0.945): Average linkage also scored very highly, forming well-balanced clusters with clear separation. Compared to single linkage, it avoids collapsing the dataset into a dominant cluster, making the results more interpretable. **Hierarchical – Complete Linkage (Silhouette = 0.909)**: Complete linkage provided a good compromise, producing balanced clusters with solid separation. The silhouette score is slightly lower than average linkage but still indicates high-quality clustering

DBSCAN (eps=0.181, min_samples=10, Silhouette = 0.470): DBSCAN detected dense clusters and successfully identified noise points. Although its silhouette score is lower than hierarchical clustering and K-Means, the method's strength lies in capturing outliers, which can reveal unusual customer behaviors.

Original Data (No Clustering): The PCA projection of the raw data shows no clear grouping, underscoring the value of clustering methods in uncovering hidden structures within the dataset.

III. Results and Evaluation

A. Performance Metrics

Table I presents the performance comparison of the clustering algorithms. Among them, hierarchical clustering with single linkage obtained the best result, achieving a silhouette score of 0.945, which reflects strong cluster separation and cohesion. In contrast, DBSCAN produced a silhouette score of 0.470, while K-means achieved 0.581.

TABLE I
Clustering Algorithm Performance

Algorithm	Silhouette Score	Intra-cluster Distance	Inter-cluster Distance
K-means	0.581	0.916	2.568
Hierarchical	0.945	1.500	28.253
DBSCAN	0.470	1.247	2.468

DBSCAN identified 6.17% of points as noise. The algorithm found 2 dense clusters among the remaining data points.

IV. Theoretical Analysis

A. Strengths and Limitations

K-Means offered moderate performance (Silhouette = 0.581). It is efficient and easy to interpret, making it suitable for business applications. However, it assumes spherical clusters and performed less effectively on irregular patterns in this dataset.

Hierarchical clustering (single and average linkage) achieved the highest silhouette scores (0.945), showing strong separation and cohesion. Its main limitation is computational cost for very large datasets and sensitivity to noise.

DBSCAN handled noise points effectively but produced the lowest silhouette score (0.470). Its usefulness lies in detecting outliers, but parameter tuning (eps, min$_{s}amples)significantlyaffectsresults.$

B. Interpretability and Business Relevance

K-Means clusters are straightforward to interpret and thus of potential value for application in customer segmentation within marketing campaigns. Hierarchical clustering provides an easy-to-understand hierarchical structure with dendrograms, which could reveal subgroup relationships. DBSCAN outlier detection provides insight into unexplained customer behavior, but the resultant clusters were less of value for business application in this data set.

C. Feature Scaling

Feature scaling was critical for this study because RFM features (Recency, Frequency, Monetary) are measured on very different scales. Monetary values can be in thousands, while frequency ranges from a few purchases, and recency spans hundreds of days. Without scaling, monetary values would dominate distance calculations. Standardization ensured that each feature contributed equally to the clustering process..

D. Stability and Sensitivity

K-Means results vary depending on initialization, which can affect stability. Hierarchical clustering is deterministic but highly sensitive to the choice of linkage method. DBSCAN's sensitivity to parameter selection made results unstable across different eps values.

E. Distance Metrics

The choice of distance metric strongly influences clustering. Euclidean distance, used in this study, works well for continuous features but is sensitive to outliers. Manhattan distance could be more robust to extreme purchase values, while cosine similarity could capture customers with similar purchase patterns regardless of scale. Mahalanobis distance, accounting for feature correlations, may be valuable given the strong correlation observed between frequency and monetary features.

V. Business Insights

The clustering analysis reveals three actionable customer segments:

High-Value Customers (Cluster 2) Action: These are the revenue spine. Retention is key—priority loyalty programs, early access to new products, and VIP customer support. Encourage advocacy by engaging them as brand ambassadors.

Active Customers (Cluster 0) Action: These customers have engagement but less consistency. They are good prospects for upselling and cross-selling. Leverage personalized product recommendations, seasonal offers, and bundles to increase both frequency and basket size.

At-Risk or Inactive Customers (Cluster 1)Action: These customers require reactivation campaigns. Use personalized win-back promotions, reminders, and limited-period discounts. Analyze past preferences to mail personalized product suggestions that generate interest.

The PCA visualization confirms that RFM features effectively capture customer behavior patterns.

VI. Conclusion

Three methods of customer clustering were compared in this project. Single linkage hierarchical clustering produced the highest silhouette score (0.945), DBSCAN (0.470) with the additional benefit of noise detection. K-means performed lower on the silhouette score (0.581), but with clusters that are easy to interpret for business use.

RFM features were proved to be effective at modeling customer behavior, and the segments derived are helpful to make knowledgeable decisions. The future course can be adding more behavioral features and trying ensemble techniques.