Customer Clustering Analysis Report

Objective: To segment customers into homogeneous groups based on transactional behaviour specifically total expenditure, frequency, and recency to facilitate the development of targeted marketing strategies.

1. Methodology

Data Integration and Feature Engineering

- **Datasets**: Clean_Customers.csv, Clean_Transactions.csv, and Clean_Products.csv were integrated using CustomerID and ProductID as primary keys.
- **Derived Features**:
  - **Total Expenditure**: Aggregate monetary value of purchases per customer.
  - **Transaction Frequency**: Total number of transactions per customer.
  - **Recency**: Days elapsed since the most recent transaction.
- **Outlier Mitigation**: Isolation Forest (contamination parameter: 2%) identified and excluded anomalous observations, ensuring robustness in clustering.
- **Preprocessing**:
  - **Logarithmic Transformation**: Applied to address right-skewed distributions in feature variables.
  - **Robust Scaling**: Features were standardized using RobustScaler to minimize outlier influence during clustering.

2. Cluster Optimization

Evaluation Criteria

- **Davies-Bouldin Index (DBI)**: Quantifies inter-cluster separation and intra-cluster cohesion (lower values denote superior separation). Target threshold: DBI < 1.
- **Silhouette Coefficient**: Measures consistency in cluster assignment (range: -1 to 1; higher values indicate better-defined clusters).

Results

- **Final DBI**: **0.8516** (satisfies target threshold).
- **Final Silhouette Score**: **0.348** (moderate cohesion).

Selection of Optimal Clusters

- **Elbow Method Analysis**: Evaluated cluster quality across 2–10 clusters.
  - **DBI Minimization**: Observed pronounced reduction in DBI up to **k=7**, beyond which marginal improvements plateaued.

- **Silhouette Score Trend**: Peak performance at **k=7**, with subsequent iterations yielding negligible gains.

- **Visual Validation**: Principal Component Analysis (PCA) projected clusters into a 2D subspace (PCA1 and PCA2), revealing distinct spatial separation with minimal overlap.

3. Cluster Profiling

Segmentation revealed seven distinct customer cohorts, differentiated by transactional behavior:

1. **High-Value Customers (Cluster 0)**:
    - Elevated total expenditure, frequent transactions, and low recency.
    - **Implication**: Prime candidates for premium loyalty programs.
2. **At-Risk Customers (Cluster 4)**:
    - Moderate historical spending but prolonged inactivity (high recency).
    - **Implication**: Require reactivation campaigns to mitigate churn.
3. **Emerging Spenders (Cluster 2)**:
    - Recent transactional activity but low frequency.
    - **Implication**: Cross-selling opportunities to boost engagement.
4. **Budget-Conscious Shoppers (Cluster 5)**:
    - Low expenditure and infrequent transactions.
    - **Implication**: Target with value-based promotions.
5. **Seasonal Shoppers (Cluster 6)**:
    - Intermittent high spending coupled with variable recency.
    - **Implication**: Time-bound offers aligned with purchase cycles.

4. Validation and Visualization

- **Dimensionality Reduction**: PCA retained 73.8% of variance in the first two components.
- **Cluster Separation**: Scatterplot visualization confirmed distinct boundaries for Clusters 0, 1, and 2, with minor overlap observed in Clusters 3 and 4, attributable to nuanced behavioral similarities.

5. Analytical Insights

- **Statistical Validation**:
    - DBI of **0.8516** signifies statistically significant differentiation between clusters.

    - Silhouette Score of **0.348** reflects moderate intra-cluster cohesion, consistent with real-world transactional heterogeneity.
- **Feature Influence**:
    - **Recency** and **Total Expenditure** emerged as primary discriminators.
    - **Transaction Frequency** further delineated high-engagement cohorts.

6. Strategic Recommendations

- **High-Value Retention**: Implement tiered rewards for Clusters 0 and 1 to reinforce loyalty.
- **At-Risk Reactivation**: Deploy personalized discounts with urgency messaging for Cluster 4.
- **Inventory Optimization**: Align stock levels with purchase patterns of dominant clusters (e.g., Cluster 0's frequent purchases).

7. Limitations and Future Directions

- **Limitations**:
    - **Silhouette Score**: Indicates opportunities for refinement in intra-cluster cohesion.
    - **Dimensionality Reduction**: PCA visualization simplifies high-dimensional data, potentially obscuring subtler patterns.
- **Future Work**:
    - Incorporate demographic variables (e.g., age, location) to enrich segmentation.
    - Evaluate alternative algorithms (e.g., DBSCAN, hierarchical clustering) for comparative analysis.

Conclusion: The K-means model with **k=7** clusters achieve statistically robust segmentation, balancing computational efficiency and interpretability. This framework provides actionable insights for personalized marketing, with avenues for enhancement through additional data integration and algorithmic exploration.

Prepared by: Maqbool Husain Saiyed

Date: 26/01/2025