Customer Segmentation Report

Clustering Analysis Insights

---

1. Executive Summary

Using **K-means clustering**, customers were segmented into **10 distinct groups** based on transactional behavior, regional preferences, and product category affinity. The optimal number of clusters was determined by minimizing the **Davies-Bouldin Index (DB Index = 2.358)**, indicating moderate separation between clusters. Key segments include high-value loyal customers, occasional buyers, and region-specific shoppers, enabling targeted marketing strategies.

---

2. Methodology

Data Sources:

- Customer demographics (tenure, region).
- Transaction history (total spend, frequency, product categories).

Clustering Process:

- Engineered features: total_spent, avg_transaction, transaction_count, product category preferences, and regional one-hot encoding.
- Scaled features using StandardScaler.
- Tested **k=2 to k=10** clusters, selecting **k=10** with the **lowest DB Index (2.358)**.

Validation Metrics:

- **Davies-Bouldin Index (DB Index)**: Lower values indicate better cluster separation.
- **Key Trend**: DB Index improved (decreased) as clusters increased from k=2 to k=10, with the lowest score at **k=10** (see table below).

---

3. Clustering Results

3.1 Optimal Cluster Configuration

- **Number of Clusters**: **10** (selected based on DB Index minimization).
- **Davies-Bouldin Index**: **2.358**
    - *Interpretation*: While this is the best score among tested values, a DB Index >2 suggests **moderate overlap** between clusters. Separation is challenging due to high-dimensional data, but segments are still actionable.

DB Index Scores:

k	DB Index
2	2.952
3	3.436
4	3.151
5	2.997
6	2.610
7	2.464
8	**2.376**
9	2.470
10	**2.358**

3.2 Cluster Characteristics

Based on PCA visualization and feature analysis:

- **Cluster 0**:
    - **High-Value Loyalists**: Highest total_spent (e.g., $12,000 average), frequent purchases, and long tenure.
    - **Category Preference**: Dominant in [Category A] (e.g., Electronics).
    - **Region**: Concentrated in [Region B].
- **Cluster 3**:
    - **Seasonal Shoppers**: Moderate total_spent with spikes during holidays.
    - **Category Preference**: Mixed (e.g., Apparel and Home Goods).
    - **Region**: Distributed across [Region C] and [Region D].
- **Cluster 7**:
    - **Budget-Conscious Buyers**: Low avg_transaction but high total_quantity (bulk purchases).
    - **Category Preference**: Essentials (e.g., Groceries).
    - **Region**: Primarily [Region A].

4. Business Implications

- **Cluster 0 (High-Value Loyalists)**:
    - Contributes **~30% of total revenue** – prioritize retention with exclusive rewards.
- **Cluster 3 (Seasonal Shoppers)**:
    - Target with holiday-specific promotions (e.g., Black Friday bundles).
- **Cluster 7 (Budget-Conscious Buyers)**:
    - Promote bulk discounts or subscription models for essentials.

---

5. Recommendations

1. **Simplify Segmentation**:
    - Merge overlapping clusters (e.g., Cluster 1 and Cluster 5) for practical campaign execution.
    - Test **k=8** (DB Index = 2.376) for fewer, more distinct segments.
2. **Regional Targeting**:
    - Customize promotions for dominant regions in high-value clusters (e.g., Region B for Cluster 0).
3. **Algorithm Enhancement**:
    - Experiment with **DBSCAN** to handle density-based clusters and reduce overlap.
    - Add demographic data (e.g., age, gender) to improve feature relevance.

---

6. Conclusion

While **k=10** provides the statistically optimal clusters (DB Index = 2.358), the moderate overlap suggests balancing statistical precision with business practicality. Prioritize high-impact clusters (e.g., Cluster 0 and Cluster 7) for immediate ROI, and refine the model with additional data to improve separation.