Technical Report: MegaMart Customer Segmentation Analysis

1. Introduction

This report summarizes the methodology, data analysis, clustering approach, validation, and results from the notebook **cluster_analysis.ipynb**, which performs **unsupervised customer segmentation** for MegaMart using **Hierarchical Clustering** and **K-Means** on behavioral and transactional variables.

The goal is to identify **natural customer segments** that support **targeted marketing, retention strategies, and resource allocation**, complementing the business-focused executive summary and strategic recommendations already defined for MegaMart's customer base.

2. Data Overview

The dataset used in cluster_analysis.ipynb contains:

- **3,000 rows** – each representing a distinct customer.
- **10 columns**, of which:
    - 1 is an identifier: customer_id.
    - **9 are behavioral/transactional variables**:
        - monthly_transactions – monthly purchase frequency.
        - avg_basket_size – average number of items per purchase.
        - total_spend – cumulative monetary spend.
        - avg_session_duration – average browsing/session time.
        - email_open_rate – email engagement ratio.
        - product_views_per_visit – depth of browsing behavior.
        - return_rate – proportion of purchases returned.
        - customer_tenure_months – time as a customer.
        - recency_days – days since last purchase.

Initial inspection from the notebook:

- There are **no missing values** in any of the variables.
- All 9 features are **numeric and continuous or ratio-scale**, suitable for distance-based clustering.
- The sample size (3,000 customers) is adequate for stable estimation of centroids and cluster profiles.

3. Exploratory Data Analysis (EDA)

EDA in the notebook includes **histograms**, a **correlation matrix**, and **scatterplots** of key variables:

- **Univariate distributions**
  - total_spend, monthly_transactions, and avg_basket_size show **right-skewed distributions**, with many low-value customers and a smaller number of high-value, heavy users.
  - return_rate is concentrated near low values, with a tail of customers who return a large fraction of their purchases.
  - recency_days reflects a mix of very recent and relatively inactive customers.
- **Correlation structure**
  From the correlation heatmap:
  - monthly_transactions, avg_basket_size, and total_spend are **strongly positively correlated**: more purchases and larger baskets are associated with higher total spend.
  - email_open_rate and product_views_per_visit are positively related to total_spend, indicating that **digital engagement tends to accompany higher value**.
  - recency_days tends to be **negatively related** to activity variables (more recent customers are typically more active).
- **Bivariate relationships**

- Scatterplots such as **Total Spend vs Monthly Transactions** and **Total Spend vs Average Basket Size** confirm that **high-spend customers either buy frequently, buy large baskets, or both**.
- No extreme outliers dominate the plots, suggesting that clustering will be driven by genuine behavioral patterns rather than a few anomalous points.

Overall, EDA supports the hypothesis that **meaningful behavioral clusters should exist**, driven mainly by shopping frequency, basket size, spending, engagement, and returns.

4. Data Preprocessing

To prepare the data for clustering, the notebook applies the following steps:

1. **Feature selection**
   - The identifier customer_id is excluded from the analysis.
   - The clustering is performed on the 9 behavioral variables stored in df_features.
2. **Standardization**
   - All features are standardized using StandardScaler to produce X_scaled.
   - This is essential because clustering algorithms such as **K-Means** and many distance-based metrics are sensitive to scale: variables with larger numerical ranges (e.g., total_spend) would otherwise dominate the distance computations compared to bounded variables (e.g., email_open_rate).

After this step, all features have **mean ≈ 0** and **standard deviation ≈ 1**, ensuring that no single variable unfairly dominates the clustering.

5. Clustering Methodology

The notebook implements a **two-stage approach**:

1. **Hierarchical clustering (Ward linkage)** to **explore structure** and identify plausible numbers of clusters.
2. **K-Means clustering** for **final partitioning** and interpretation, supported by quantitative validation (inertia and silhouette scores).

5.1 Hierarchical Clustering

- The standardized data X_scaled is used to compute hierarchical linkages with methods:
    - "single", "complete", "average", and "ward".
- For each method, dendrograms are plotted to visually inspect cluster merging behavior.

The **Ward linkage** method is chosen as the primary reference because it:

- Minimizes **within-cluster variance** at each merge.
- Produces **compact, well-separated clusters** that are easier to interpret.

Using Ward linkage, the notebook:

- Draws a dendrogram with a horizontal cut at a chosen distance (e.g., ~70 units).
- Evaluates **candidate numbers of clusters** $k = 3,4,5,6$using fcluster and **silhouette scores**, obtaining:
- $k = 3$: silhouette ≈ **0.2948**
- $k = 4$: silhouette ≈ **0.3157**
- $k = 5$: silhouette ≈ **0.3003**
- $k = 6$: silhouette ≈ **0.2475**

This analysis suggests that $k = 4$provides a **good trade-off** between cluster separation and model complexity, with the highest silhouette among the tested values.

5.2 K-Means Clustering

On the standardized data, the notebook runs **K-Means** for $k = 2,3,\ldots,10$with n_init = 10, recording **inertia** and **silhouette score**:

- Example results:
    - $k = 2$: inertia = 19065.45, silhouette ≈ **0.3446**
    - $k = 3$: inertia = 14397.53, silhouette ≈ **0.2974**
    - $k = 4$: inertia = 11944.28, silhouette ≈ **0.3173**
    - $k = 5$: inertia = 10616.55, silhouette ≈ **0.2696**

  - ... up to $k = 10$, with progressively lower inertia and mixed silhouette scores.

Key observations:

- The **silhouette** is highest at $k = 2$but remains **competitive at** $k = 4$.
- The **inertia curve** (Elbow Method) shows a marked reduction from $k = 2$to $k = 4$, after which the curve **starts to flatten**, indicating diminishing returns.

Given the **hierarchical results** and **K-Means validation**, the notebook selects:

Final number of clusters: $k = 4$

This choice balances:

- Model interpretability and business usefulness.
- Adequate separation of customer behaviors.
- The ability to distinguish more nuanced segments than a simple 2-cluster solution.

6. Model Evaluation & Validation

6.1 Silhouette Analysis

For the final K-Means solution with $k = 4$:

- The **average silhouette score** is ≈ **0.3173**, indicating **moderate** cluster separation.
- A detailed silhouette plot per cluster shows that:
  - **Clusters 0 and 3** contain predominantly **positive and relatively high silhouette values**, indicating **well-defined clusters**.
  - **Clusters 1 and 2** exhibit a **small fraction of negative silhouette values**, indicating **some customers are borderline** and could plausibly belong to another cluster.

Despite these borderline cases, the overall silhouette structure **supports the choice of 4 clusters** as a reasonable segmentation.

6.2 PCA Visualization

To visualize cluster separation in a reduced space, the notebook applies **Principal Component Analysis (PCA)** to X_scaled:

- **PC1 + PC2 explain ≈ 61.98% of total variance**

  - PC1 ≈ 41.01%
  - PC2 ≈ 20.97%

A scatterplot of **PC1 vs PC2**, colored by K-Means cluster labels and showing centroids, reveals:

- Cluster 0 and Cluster 3 appear as **relatively distinct regions**.
- Cluster 1 and Cluster 2 show **some overlap** but still occupy characteristic zones in the PCA plane.

Because PCA compresses a **9-dimensional space into 2 dimensions**, the separation is an approximation, but it visually confirms the **moderate but meaningful structure** indicated by the silhouette metrics.

7. Cluster Profiling & Segment Interpretation

After fitting K-Means with $k = 4$, the notebook:

- Stores the labels in df["cluster_kmeans"].
- Computes **cluster sizes and percentages**:

Cluster	Count	Percentage
0	525	17.50%
1	929	30.97%
2	433	14.43%
3	1113	37.10%

It then calculates **mean values per feature and cluster** (original and normalized). The normalized (z-score) profile highlights how each cluster differs from the global average:

- Positive z-score: above-average value.
- Negative z-score: below-average value.

Below is a summary of each cluster's behavioral profile and business interpretation, aligned with the segments described in the executive summary.

Cluster 0 (17.50%) – Loyal Customers / High-Value Engaged

Data profile (original means):

- monthly_transactions: 14.07 (highest)
- avg_basket_size: 22.03 (highest)
- total_spend: 6507.29 (highest)
- avg_session_duration: 45.92 (long)
- email_open_rate: 0.576 (highest)
- product_views_per_visit: 43.01 (highest)
- return_rate: 0.099 (lowest)
- customer_tenure_months: 26.22 (longest)
- recency_days: 8.02 (most recent)

Interpretation:

- These customers **buy frequently**, with **large baskets** and **very high total spend**.
- They are **highly engaged digitally**, opening emails and viewing many products per visit.
- They **rarely return items**, suggesting trust and satisfaction.
- Long tenure and low recency confirm that they are **long-standing, actively loyal clients**.

Segment label: Loyal Customers – core revenue drivers and prime candidates for **VIP programs, exclusive offers, and personalized experiences**.

Cluster 1 (30.97%) – High Returners / Hesitant Explorers

Data profile:

- monthly_transactions: 1.68 (very low)
- avg_basket_size: 3.05 (very low)
- total_spend: 422.62 (lowest)
- avg_session_duration: 52.31 (longest sessions)

- email_open_rate: 0.374 (lowest engagement)
- product_views_per_visit: 30.01 (moderate browsing)
- return_rate: 0.275 (highest)
- customer_tenure_months: 15.31 (moderate/low)
- recency_days: 35.59 (least recent)

Interpretation:

- This group **browses a lot (long sessions)** but makes **few and small purchases**.
- They have the **highest return rate**, indicating **lack of confidence, mismatch of expectations, or issues with product information**.
- Low email engagement and high recency suggest they are **drifting away** or never fully committed to the brand.

Segment label: High Returners – curious but unconvinced customers who require **clear product information, reassurance, and strong conversion incentives** (e.g., free returns, live chat support, detailed reviews).

Cluster 2 (14.43%) – Seasonal Shoppers / High-Value Infrequent

Data profile:

- monthly_transactions: 4.04 (low-moderate)
- avg_basket_size: 18.17 (very high)
- total_spend: 3875.94 (high)
- avg_session_duration: 22.36 (short)
- email_open_rate: 0.450 (moderate)
- product_views_per_visit: 16.55 (lower than average)
- return_rate: 0.245 (high)
- customer_tenure_months: 21.60 (medium-long)
- recency_days: 19.84 (moderate)

Interpretation:

- These customers purchase **infrequently**, but when they do, they create **large baskets and high total spend**.
- Their browsing is **more focused** (shorter sessions, fewer product views), but they have **above-average return rates**.
- This behavior is consistent with **event-driven or seasonal buying** (e.g., holidays, back-to-school, large household purchases), sometimes leading to returns when large orders contain mismatched or unnecessary items.

Segment label: Seasonal Shoppers – important but sporadic customers who respond well to **timely, event-based campaigns and bundle offers** aligned with key calendar moments.

Cluster 3 (37.10%) – New and Casual Shoppers

Data profile:

- monthly_transactions: 6.59 (moderate)
- avg_basket_size: 5.56 (small)
- total_spend: 1450.95 (low-moderate)
- avg_session_duration: 29.60 (medium)
- email_open_rate: 0.437 (moderate)
- product_views_per_visit: 32.75 (moderate-high)
- return_rate: 0.130 (low)
- customer_tenure_months: 14.73 (shortest)
- recency_days: 14.53 (relatively recent)

Interpretation:

- This is the **largest cluster** by far.
- Customers shop at a **moderate frequency**, but with **small baskets and relatively low spend**.
- They show **moderate engagement** (email and product views), **low returns**, and **short tenure**, suggesting they are **newer customers still exploring the brand**.

Segment label: New and Casual Shoppers – a broad base of low-to-medium value customers who can be nurtured via **onboarding journeys, recommendations, and introductory offers** to increase loyalty and spend.

8. Discussion

From a technical perspective, the clustering workflow in cluster_analysis.ipynb is coherent and well justified:

- **Preprocessing** (feature selection and standardization) ensures comparable scales.
- **Hierarchical clustering** provides an exploratory view of structure and supports the choice of $k = 4$based on silhouette scores.
- **K-Means** offers a compact, centroid-based segmentation that is easy to deploy operationally (e.g., assigning segments in a CRM).
- **Validation** via inertia, silhouette scores, and PCA visualization confirms that clusters are **moderately well separated**, with especially clear structure for high-value loyal customers and the large base of new/casual shoppers.

From a business standpoint, the four clusters match the segments described in the executive summary: **Loyal Customers, High Returners, Seasonal Shoppers, and New & Casual Shoppers**, each requiring different engagement and retention strategies.

Some limitations noted from the analysis:

- The **average silhouette (~0.32)** indicates **moderate**, not perfect, separation; some overlap is expected for complex real behaviors.
- **Clusters 1 and 2** show a few negative silhouette cases, highlighting customers whose behavior is ambiguous between segments.
- Only **behavioral variables** are used; enriching the model with demographic or product-category preferences could further refine segments.

Despite these limitations, the solution provides **actionable segmentation** that aligns well with intuitive marketing personas and can be directly integrated into MegaMart’s decision-making.

9. Conclusion

The cluster_analysis.ipynb notebook successfully demonstrates the application of **multivariate methods and clustering** for customer segmentation in MegaMart:

- A dataset of **3,000 customers and 9 behavioral features** was cleaned, explored, and standardized.
- **Hierarchical clustering** (Ward linkage) and **K-Means** were used in combination to determine an appropriate number of clusters.
- Based on elbow and silhouette criteria, **four clusters** were selected as the **optimal solution**.
- Detailed cluster profiling produced **four meaningful customer segments**:
    1. **Loyal Customers** – high value, highly engaged, low returns.
    2. **High Returners** – low spend, high returns, long browsing, low commitment.
    3. **Seasonal Shoppers** – infrequent but high-value purchases, often event-driven.
    4. **New and Casual Shoppers** – large group of relatively new, low-to-medium value customers.

These results align with and technically ground the executive summary, offering a clear bridge from **data science outputs to marketing strategy**. In line with the format of the LendSmart report, this document consolidates the notebook's methodology, validation, and conclusions into a concise technical narrative suitable for academic and business audiences