

## Clustering Methodology and Concept

### # Methodology

#### 1. Data Preparation:

- ◆ Merged customer profiles and transaction data for a comprehensive view.
- ◆ Aggregated transactional data (Price, Category) per customer.
- ◆ Encoded categorical variables (Product Categories) using one-hot encoding.
- ◆ Normalized numerical features to ensure equal contribution to clustering.

#### 2. Clustering Algorithm:

- ◆ **KMeans Clustering** was used due to its simplicity and efficiency for large datasets.
- ◆ Tested cluster counts (k) between 2 and 10 to find the optimal number of clusters.
- ◆ Used Davies-Bouldin (DB) Index to evaluate clustering quality.

#### 3. Evaluation Metrics:

- ◆ **Davies-Bouldin Index:** Measures the compactness and separation of clusters; lower values indicate better clustering.
- ◆ **Silhouette Score:** Evaluates how well each point fits within its cluster.

#### 4. Dimensionality Reduction:

- ◆ Applied PCA to reduce high-dimensional data for better visualization and interpretation.

### # Results

#### 1. Optimal Clusters:

- ◆ Determined the optimal number of clusters based on the lowest DB Index value.
- ◆ *5 clusters were found to provide the best segmentation.*

#### 2. Cluster Insights:

- ◆ **Cluster 0:** High spenders with frequent purchases in premium categories.
- ◆ **Cluster 1:** Occasional buyers with moderate spending habits.
- ◆ **Cluster 2:** Frequent low-cost purchases across diverse categories.
- ◆ **Cluster 3:** Rare but high-value transactions.

#### 3. Metrics:

- ◆ Davies-Bouldin Index for Optimal Clusters: 0.9252322177528702
- ◆ Silhouette Score: 0.02863924368427505
- ◆ Number of Clusters: 5

## # Visualization

- Clusters were visualized using a 2D scatterplot (PCA-reduced data).
- Clear separation of clusters highlighted distinct customer segments.

