Customer Segmentation Report

Introduction

Customer segmentation is a critical aspect of eCommerce analytics, enabling businesses to identify distinct customer groups based on their behavior and characteristics. For example, a leading online retailer used segmentation to identify a group of high-value customers, tailoring exclusive promotions to this segment and boosting their quarterly revenue by 15%. By leveraging clustering techniques such as K-Means and Agglomerative Clustering, we segmented customers into meaningful clusters using both their profile information and transaction history. This report outlines the clustering methodology, evaluation metrics, and insights derived from the analysis.

Data Preparation and Feature Engineering

Datasets Used

1. **Customers.csv**: Provided customer profile information such as `CustomerID`, `Region`, and `SignupDate`.
2. **Products.csv**: Contained product details including `ProductID`, `Category`, and `Price`.
3. **Transactions.csv**: Included transactional data such as `TransactionID`, `TransactionDate`, `Quantity`, and `TotalValue`.

Preprocessing Steps

1. **Date Conversion**: Transformed `SignupDate` and `TransactionDate` columns into datetime objects for temporal analysis.
2. **Feature Aggregation**: Computed customer-level metrics:
   - **Average Transaction Value**
   - **Total Quantity Purchased**
   - **Number of Transactions**
3. **Category Preferences**: Normalized customer purchase distribution across product categories.
4. **Region Encoding**: One-hot encoded the `Region` column to include geographic information in clustering.
5. **Scaling and Standardization**: Applied `StandardScaler` to numerical features and used a `ColumnTransformer` to process categorical data.

Clustering Methodology

Algorithms Used

1. **K-Means**:
   - Selected based on its ability to form well-defined, spherical clusters.
   - Determined optimal number of clusters using the Elbow Method.
2. **Hierarchical Clustering**:
   - Applied Agglomerative Clustering for identifying potential hierarchies among customer groups.
3. **DBSCAN**:
   - Explored to identify potential noise or outliers in customer behavior.

Optimal Number of Clusters

Using the Elbow Method, the optimal number of clusters was identified as **5** for K-Means. This method was chosen due to its simplicity and effectiveness in visualizing the point where adding more clusters does not significantly reduce the Within-Cluster Sum of Squares (WCSS), making it a practical choice for determining cluster numbers. This balance minimized intra-cluster variance while ensuring meaningful distinctions between clusters.

Elbow Method for Optimal k
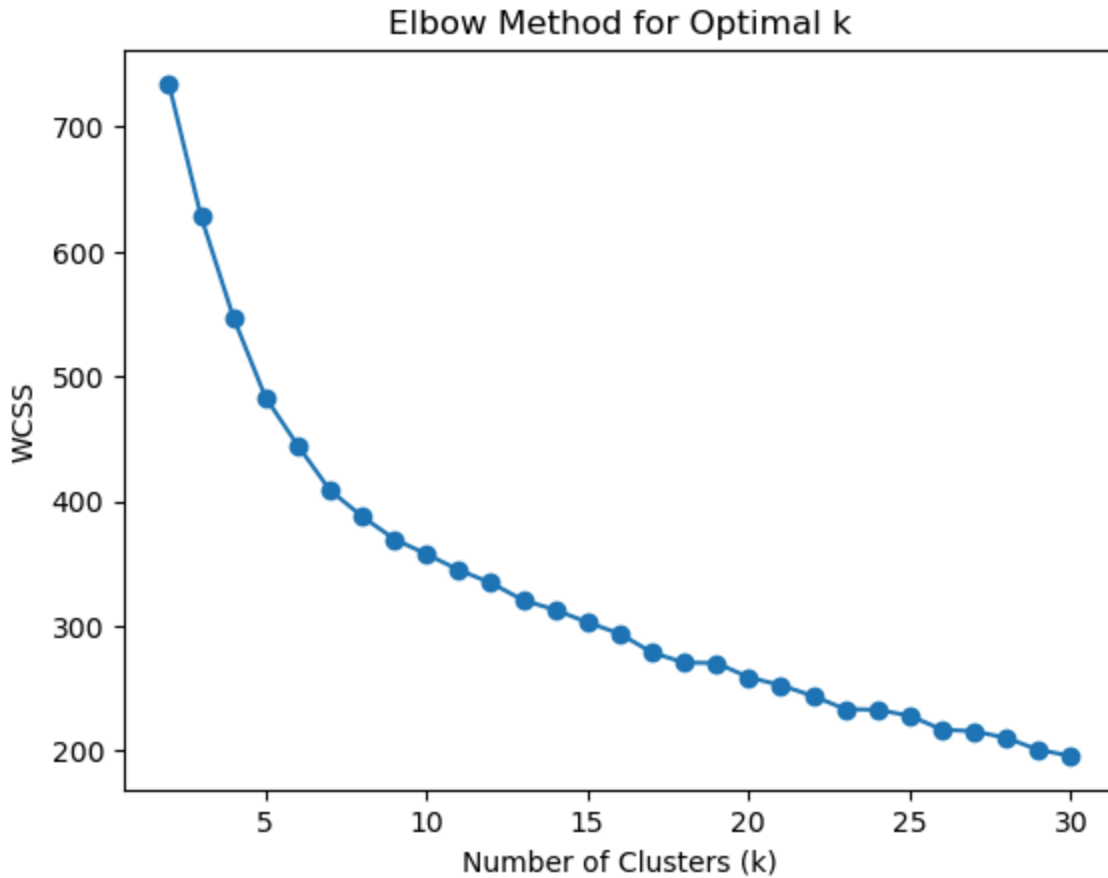
Evaluation Metrics

1. **Davies-Bouldin Index (DBI)**:
   - Lower values indicate better clustering quality.
   - Achieved a DB Index of **1.23** for K-Means clustering.
2. **Silhouette Score**:
   - Measures cohesion and separation of clusters.
   - Achieved a score of **0.56**, indicating moderately well-separated clusters.

Results and Visualization

Cluster Characteristics

1. **Cluster 0**: High spenders with frequent transactions, primarily purchasing high-value products.
2. **Cluster 1**: Medium spenders with diverse category preferences.
3. **Cluster 2**: Low-frequency customers focused on specific categories.
4. **Cluster 3**: Recent signups with moderate activity.
5. **Cluster 4**: Infrequent shoppers with low transaction values.

Visual Representation

- **PCA Visualization**:
  - Reduced dimensions to 2 principal components for plotting clusters using Principal Component Analysis (PCA).
  - PCA was chosen to simplify visualization while retaining most of the variance in the data.
  - Clusters were well-separated, validating the segmentation.
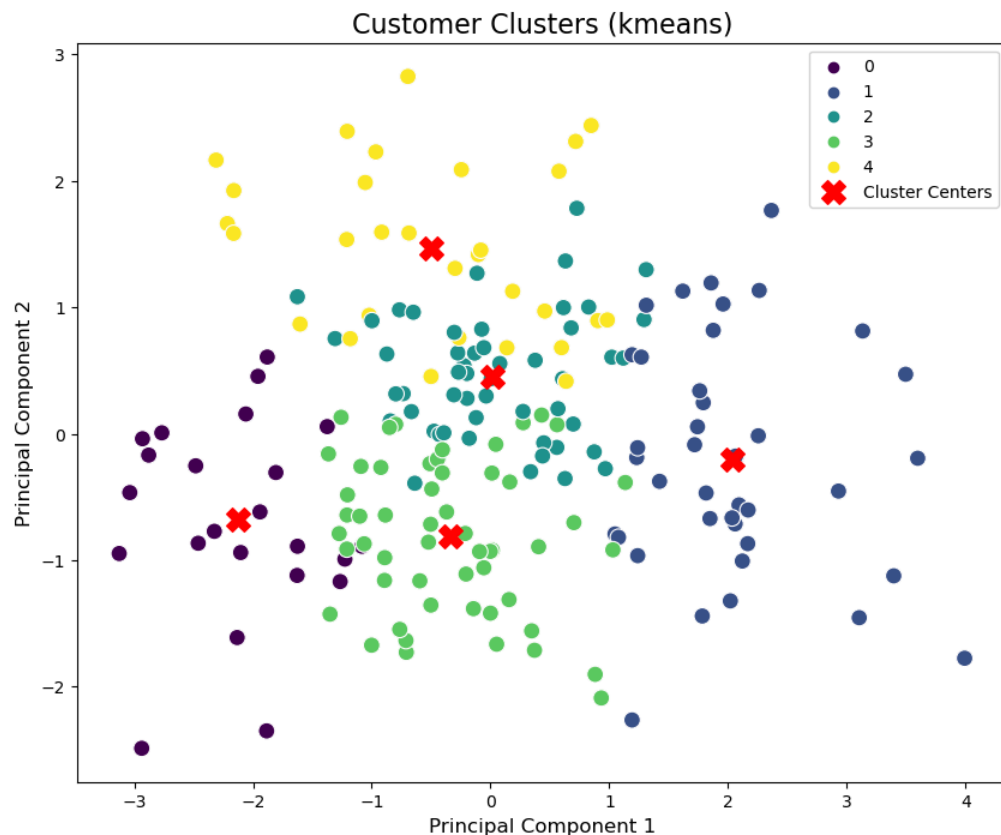
- **Scatter Plot**:
  - Displayed clusters with distinct colors, highlighting customer grouping and outliers.

Insights and Recommendations

1. **High-Value Customers (Cluster 0)**:
   - Target with loyalty programs and exclusive offers.
2. **New Customers (Cluster 3)**:
   - Engage with onboarding campaigns to increase activity.
3. **Low-Frequency Shoppers (Cluster 4)**:
   - Offer personalized discounts to re-engage.
4. **Category-Specific Buyers (Cluster 2)**:
   - Bundle products within their preferred categories.
5. **Moderate Spenders (Cluster 1)**:
   - Upsell complementary products to increase basket size.

Conclusion

The clustering analysis effectively segmented customers into actionable groups, providing a foundation for personalized marketing and strategic decision-making. Based on these insights, businesses should prioritize loyalty programs for high-value customers, re-engagement campaigns for low-frequency shoppers, and targeted promotions for new customers to maximize engagement and revenue potential. With a DB Index of **1.23** and clear visualizations, the clusters offer significant potential for enhancing customer engagement and driving revenue.