Customer Segmentation Report

Clustering Algorithm: K-Means **Number of Clusters (k):** 6

1. Introduction

This report presents the results of customer segmentation performed on the eCommerce transactions dataset. The K-Means clustering algorithm was used to group customers into distinct segments based on their purchasing behavior and profile characteristics. The goal of this analysis is to identify meaningful customer segments that can inform targeted marketing strategies and improve customer engagement.

2. Methodology

The following steps were taken to perform the customer segmentation:

1. **Data Preprocessing:**
   - Loaded the `Customers.csv`, `Products.csv`, and `Transactions.csv` datasets.
   - Converted date columns to datetime objects.
   - Merged transaction data with product data to obtain product category information.
2. **Feature Engineering:**
   - **Recency:** Calculated the number of days since the customer's last purchase.
   - **Frequency:** Calculated the total number of purchases made by each customer.
   - **Monetary Value:** Calculated the total amount spent by each customer.
   - **Product Category Preferences:** Created binary features (one-hot encoding) for each product category, indicating whether a customer had purchased from that category.
   - **Signup Year and Month:** Extracted the year and month of customer signup.
   - **Region:** One-hot encoded the customer's region.
3. **Handling Missing Values:**
   - Filled missing values in `Recency` with the maximum recency value (indicating no prior purchase).
   - Filled missing values in `Frequency` and `MonetaryValue` with 0.
   - Filled missing values in all other columns (one hot encoded category and region columns) with 0.
4. **Feature Scaling:**
   - Standardized numerical features using `StandardScaler` to ensure that features with larger values did not disproportionately influence the distance calculations.
5. **Clustering:**
   - Applied the K-Means algorithm with `k=6` clusters, as determined by the analysis of the Elbow Method, Davies-Bouldin Index, Silhouette Score, and Calinski-Harabasz score.
6. **Evaluation and Visualization:**
   - Calculated clustering metrics: Davies-Bouldin Index, Silhouette Score, Calinski-Harabasz Score.
   - Generated scatter plots to visualize the clusters in a 2D space using different feature combinations.
   - Analyzed the mean values of features within each cluster to understand their characteristics.

3. Clustering Results

The K-Means algorithm with `k=6` produced the following clustering metrics:

- **Davies-Bouldin Index:** 1.8392353682316722
- **Silhouette Score:** 0.12766301976648628
- **Calinski-Harabasz Score:** 27.5254260358545

Interpretation of Metrics:

- **Davies-Bouldin Index (DBI):** A lower DBI indicates better clustering. The value suggests that the clusters are relatively well-separated, but there might be some overlap.
- **Silhouette Score:** The score ranges from -1 to +1, with higher values indicating better-defined clusters. The score suggests that the clusters are somewhat defined but there's room for improvement in the cluster structure.
- **Calinski-Harabasz Score:** A higher Calinski-Harabasz score indicates better-defined clusters. The score suggests that the clusters are relatively well-defined.

4. Conclusion

The K-Means clustering algorithm with `k=6` identified six distinct customer segments based on the provided dataset. These segments exhibit varying levels of recency, frequency, monetary value, and product category preferences. The Davies-Bouldin

Index, Silhouette Score, and Calinski-Harabasz score suggest that the clusters are relatively well-defined, although there is some room for improvement in cluster separation.