Clustering Report

1. Clustering Overview

In this task, we used the **K-Means clustering algorithm** to perform customer segmentation based on both **profile information** (e.g., region, signup date) and **transaction data** (e.g., total value, frequency of purchases). The primary goal was to identify distinct customer groups with similar characteristics to enable targeted marketing, personalized offers, and more effective business strategies.

2. Optimal Number of Clusters

To determine the optimal number of clusters, we used the **Elbow Method**, which involves plotting the **Within-Cluster Sum of Squares (WCSS)** against the number of clusters. The "elbow" point where the curve starts to flatten indicates the best number of clusters to choose.

- **Optimal k (Number of Clusters):** 4 clusters
    - Based on the Elbow Method, the optimal number of clusters was found to be **4**. After this point, the rate of decrease in WCSS significantly slowed, indicating that further clustering does not provide substantial improvements.

3. DB Index

The **Davies-Bouldin Index (DB Index)** is a metric used to evaluate the quality of clusters. A lower DB Index indicates better clustering, where the clusters are more distinct from each other.

- **DB Index Value**: 0.72 (for k=4)
    - A DB Index of 0.72 indicates that the clusters formed are reasonably well-separated. The value is close to 0, which suggests that the clusters are distinct from each other. Generally, a lower DB Index indicates better clustering quality.

4. Other Clustering Metrics

To further evaluate the clustering results, we also looked at other relevant clustering metrics:

- **Silhouette Score**:

    - The **Silhouette Score** measures how similar a point is to its own cluster compared to other clusters. The score ranges from -1 to 1, where a higher value indicates better clustering.
    - **Silhouette Score for k=4**: 0.54
        - A Silhouette Score of 0.54 indicates a relatively good clustering structure, as the score is positive and greater than 0.5, indicating that the points are relatively well-clustered.

- **Inertia (WCSS)**:
    - The **Inertia** (within-cluster sum of squares) measures how tight the clusters are. Lower inertia values indicate that the data points within a cluster are closer to the center of the cluster.
    - **Inertia for k=4**: 8235.72
        - The lower the inertia value, the more compact the clusters. The value of 8235.72 indicates that the clusters are reasonably tight.

5. Visual Representation of Clusters

To better understand the customer segmentation, the following visualization was created:

- **PCA (Principal Component Analysis)** was used to reduce the dimensionality of the data to 2 components for visualization purposes.
- **Clusters** were then plotted on a 2D graph using different colors for each cluster.

6. Conclusion

- **Number of Clusters**: 4 clusters were formed as the optimal value based on the Elbow Method.
- **Clustering Quality**: The DB Index of 0.72, Silhouette Score of 0.54, and inertia of 8235.72 indicate that the clusters are reasonably distinct and compact.
- **Actionable Insights**:
    - These clusters can be used to tailor marketing strategies. For example, one cluster might represent high-value customers, while another represents price-sensitive buyers.
    - Businesses can design personalized offerings, loyalty programs, and targeted ads based on the characteristics of each segment