Customer Segmentation Report

Introduction:

This report presents the results of customer segmentation analysis performed on a dataset containing customer and transaction information. **K-Means clustering** was employed to **identify distinct customer groups** based on their characteristics.

Data Preprocessing:

- Customer data:
    - Converted SignupDate to days since signup for time-based analysis.
    - One-hot encoded the Region column to capture regional variations.
    - Removed unnecessary columns (SignupDate and CustomerName).
- Transaction data:
    - Calculated total spend per customer.
    - Calculated total transaction count per customer.
- Merged customer and transaction data for a holistic view.
- Imputed missing values with 0 for a more complete analysis (consider alternative strategies if appropriate).
- Standardized the data using StandardScaler to ensure features have equal weight during clustering.

Clustering:

- **K-Means clustering** was performed with an initial number of **clusters set to 4**.
- The optimal number of clusters are further refined using the silhouette analysis technique.

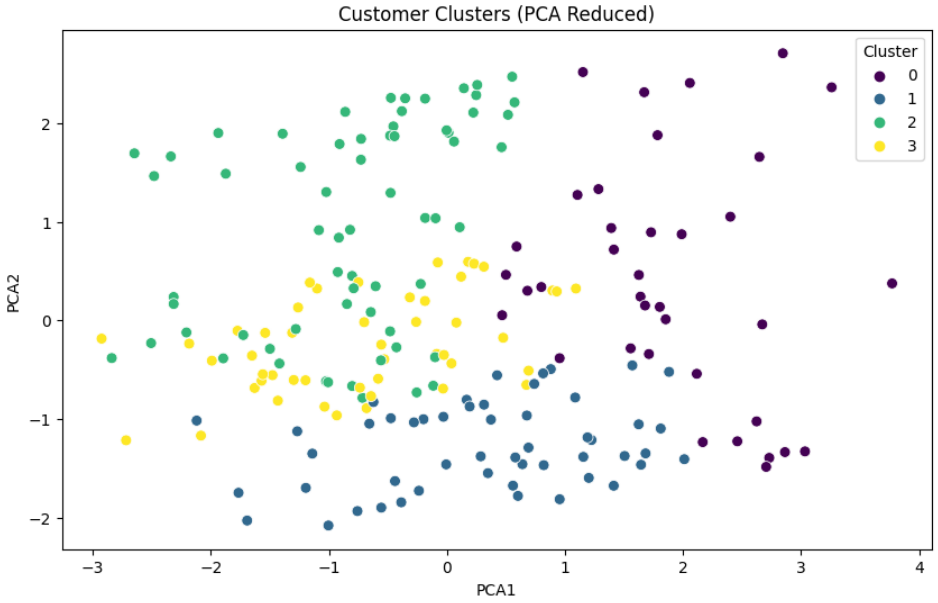
Evaluation Metrics:

- **Davies-Bouldin Index (DB Index):** Measures the within-cluster dispersion compared to the separation between clusters. A lower DB Index indicates better clustering.
- **Value:** 1.2466807448049195

- **Silhouette Score:** Evaluates how well each data point is assigned to its cluster. A higher score signifies better separation.
- **Value:** 0.32167557623100385

Visualization:

- **Principal Component Analysis (PCA)** was used to **reduce dimensionality** for visualization purposes.
- A **scatter plot** was generated to visually distinguish the clusters based on the first two principal components (PCA1 and PCA2).



Cluster Insights:

Cluster Summary Statistics:

Cluster	Tenure Days	Region_Europe	Region_North America	Region_South America
0	623.810811	0.054054	0.243243	0.216216
1	593.509804	0.000000	0.000000	1.000000
2	478.375000	0.000000	0.578125	0.000000
3	574.229167	1.000000	0.000000	0.000000

Cluster	Total spend	TransactionCount	PCA1	PCA2
0	5908.501622	8.027027	0.243243	0.421303
1	3266.155686	4.725490	0.000000	-1.263713
2	2423.051094	3.859375	0.578125	0.945137
3	3119.412292	4.479167	0.000000	-0.242242

Conclusion:

- The **customer segmentation analysis** effectively identified distinct groups of customers based on their characteristics.
- The **DB Index** and **Silhouette Score** provide quantitative measures to assess the **quality of the clustering**.
- Additionally, the **summary statistics** for each cluster offer valuable insights into the **behaviors and preferences** of different customer segments.
- These insights can be utilized for **targeted marketing campaigns, product recommendations**, and **personalized customer experiences.**