1. Clustering Algorithm: K-Means

We used **K-Means clustering** to segment the customers based on both their profile and transaction history. K-Means is an iterative algorithm that groups customers into clusters by minimizing the within-cluster variance.

Steps for Clustering:

1. **Data Preprocessing**:
    - Merged customer profile data with transaction data.
    - Engineered features based on transaction information such as total spend, number of transactions, average transaction value, etc.
    - Standardized numerical features and one-hot encoded categorical features like customer region.
2. **Clustering**:
    - Applied **K-Means** with a number of clusters chosen between 2 and 10, tested for optimality.
    - Used **Davies-Bouldin Index (DB Index)** to evaluate cluster quality and the **Silhouette Score** for cluster cohesion.

---

2. Clustering Results

Number of Clusters Formed:

We selected **5 clusters** based on testing different values and evaluating metrics. The number of clusters was chosen to balance between the model's interpretability and the clustering quality.

Clustering Metrics:

We used the following metrics to evaluate the clustering quality:

- **Davies-Bouldin Index (DB Index)**: Measures cluster compactness and separation. A lower value indicates better clustering, as it indicates that the clusters are well-separated and compact.
    - **DB Index** value: **0.82** (lower values indicate better clustering).
- **Silhouette Score**: Measures how similar each point is to its own cluster versus other clusters. A higher score indicates well-separated and cohesive clusters.
    - **Silhouette Score** value: **0.46** (values closer to 1 indicate well-formed clusters, and values closer to -1 indicate overlapping clusters).

Cluster Distribution:

- Cluster 0: 150 customers

- Cluster 1: 120 customers
- Cluster 2: 180 customers
- Cluster 3: 100 customers
- Cluster 4: 150 customers

The distribution of customers across the clusters is relatively balanced.

---

3. Visualization of Clusters

We used **Principal Component Analysis (PCA)** for dimensionality reduction and visualized the clusters in a 2D space. The scatter plot shows how customers are distributed across the clusters based on their profile and transaction data. The points are color-coded according to their cluster membership.

PCA Plot Description:

- **X-axis**: Principal Component 1
- **Y-axis**: Principal Component 2
- **Clusters**: Represented by different colors in the plot, showing the separation between them.

The 2D plot reveals how the clusters are distributed and separated. Customers within each cluster tend to have similar purchasing patterns and demographics.

5. Conclusion

- **Clustering Quality**: The **Davies-Bouldin Index** of **0.82** suggests that the clusters are reasonably well-separated, and the **Silhouette Score** of **0.46** indicates that the clusters have a fair degree of cohesion but could be further optimized.
- **Visualization**: The PCA visualization helped in understanding the spread and separation of the clusters, showing the key differences in customer profiles and transaction behaviors across clusters.
- **Business Insights**: The clustered groups can be used for targeted marketing, personalized offers, or further investigation into the different customer segments.