Project Title: Customer Segmentation for Bank's Credit Card Holders

Overview:
This project aims to develop a customer segmentation model to analyse the transaction and credit usage patterns of credit card holders for a bank. The goal is to group customers into different segments based on their behaviour, enabling the bank to devise personalised marketing strategies and improve customer satisfaction.

Objectives:
- Perform data preprocessing and exploratory data analysis (EDA) on the dataset.
- Apply clustering algorithms (K-Means, DBSCAN) to segment customers.
- Evaluate cluster quality using silhouette scores and visualize the clustering results.
- Profile each segment based on transaction patterns and suggest marketing strategies.

1. Data Preprocessing
The dataset provided contains details of credit card transactions and customer information. Several preprocessing steps were carried out to ensure the dataset is clean and ready for clustering:

Steps:
1. Dropped non-numeric columns:
The "CUST_ID" column, which serves as a unique identifier but does not contribute to clustering, was dropped.

2. Handled missing values:
Missing values in numeric columns were filled with the median of each respective column to maintain the data distribution.

3. Feature engineering:
- Created new features like:
  - `AVG_TRANSACTION_AMOUNT`: Average transaction amount per month (calculated as `PURCHASES / TENURE`).
  - `TOTAL_CREDIT_USED`: Total credit used, which is the sum of the balance and cash advance.
  - `CREDIT_USED_RATIO`: Ratio of credit used to the credit limit.
  - `TRANSACTION_FREQUENCY`: Frequency of transactions made per month.

4. Feature scaling:
StandardScaler was used to standardise numerical features, ensuring all features are on the same scale for clustering.

2. Exploratory Data Analysis (EDA)
EDA was performed to understand the distribution of data, detect patterns, and identify outliers:

- **Histograms:** A histogram was plotted for the average transaction amount to visualize the distribution.
- **Box Plot:** A box plot for credit usage was created to detect outliers in the usage ratio.
- **Correlation Heatmap**: A heatmap was generated to display correlations between various features such as credit usage, balance, payments, and purchases.

---

3. Clustering Algorithms
Two clustering algorithms, K-Means and DBSCAN(Density based spatial clustering of application with noise), were applied to segment the customers based on their transaction and credit usage patterns.

3.1 K-Means Clustering:
- **Elbow Method:** Used to determine the optimal number of clusters by evaluating the sum of squared distances (inertia) for different values of K.
- **Silhouette Scores:** Calculated to assess the quality of the clusters for different values of K (range 2–10).
- The optimal number of clusters was found based on the highest silhouette score.

3.2 DBSCAN Clustering:
- Different combinations of epsilon and minimum sample values were tested for DBSCAN clustering.
- Silhouette scores were computed to determine the best parameters for DBSCAN.

4. Cluster Profiling
After determining the optimal clusters using K-Means, the following insights were extracted for each segment:

- Average Transaction Amount
- Total Spending
- Transaction Frequency
- Credit Usage
- Balance

Each segment was profiled based on the above metrics to better understand customer behavior.

5. Marketing Strategies
Based on the profiles of the clusters, marketing strategies were proposed for each group:

- **High Spenders:** Premium rewards and exclusive offers to enhance loyalty.
- **Medium Spenders:** Targeted campaigns with personalized offers to encourage higher spending.
- **Low Transaction Frequency Customers**: Incentives for increased usage, such as cashback offers.

6. Dimensionality Reduction and Visualization

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the data for better visualization of clusters. Both K-Means and DBSCAN clustering results were visualized using the reduced dimensions from PCA.

7. Evaluation of Clustering Performance

The silhouette scores for both K-Means and DBSCAN were calculated for both the original and PCA-transformed data. K-Means performed better in this dataset, showing more distinct cluster formations.

Conclusion:

This project successfully segments credit card customers into distinct groups based on transaction and credit usage patterns. These insights can help the bank tailor marketing strategies, leading to improved customer satisfaction and retention.