**Data Science Assignment: eCommerce Transactions Dataset**

**Task 3: Customer Segmentation / Clustering**

**Code Implementation**

|  |
| --- |
| *import pandas as pd*  *import numpy as np*  *from sklearn.cluster import KMeans*  *from sklearn.preprocessing import StandardScaler*  *from sklearn.metrics import davies\_bouldin\_score, silhouette\_score*  *import matplotlib.pyplot as plt*  *import seaborn as sns*  *from sklearn.decomposition import PCA* |

|  |
| --- |
| *customers = pd.read\_csv('Customers.csv')*  *transactions = pd.read\_csv('Transactions.csv')*  *customers.fillna({'Region': 'Unknown', 'SignupDate': customers['SignupDate'].mode()[0]}, inplace=True)*  *transaction\_summary = transactions.groupby('CustomerID').agg(*  *TotalValue=('TotalValue', 'sum'),*  *TransactionCount=('TransactionID', 'count')*  *).reset\_index()*  *customer\_data = customers.merge(transaction\_summary, on='CustomerID', how='left')*  *customer\_data['Region'] = customer\_data['Region'].map({*  *'North America': 0, 'Europe': 1, 'Asia': 2, 'South America': 3, 'Africa': 4, 'Australia': 5, 'Unknown': 6*  *})*  *customer\_data['SignupDate'] = pd.to\_datetime(customer\_data['SignupDate'])*  *customer\_data['DaysSinceSignup'] = (pd.to\_datetime('today') - customer\_data['SignupDate']).dt.days*  *features = customer\_data[['Region', 'TotalValue', 'TransactionCount', 'DaysSinceSignup']]* |

|  |
| --- |
| *scaler = StandardScaler()*  *scaled\_features = scaler.fit\_transform(features)* |

|  |
| --- |
| *kmeans = KMeans(n\_clusters=5, random\_state=42) # You can experiment with different k*  *customer\_data['Cluster'] = kmeans.fit\_predict(scaled\_features)*  *db\_index = davies\_bouldin\_score(scaled\_features, customer\_data['Cluster'])*  *silhouette = silhouette\_score(scaled\_features, customer\_data['Cluster'])*  *print(f'Davies-Bouldin Index: {db\_index}')*  *print(f'Silhouette Score: {silhouette}')* |

|  |
| --- |
| *pca = PCA(n\_components=2)*  *pca\_components = pca.fit\_transform(scaled\_features)*  *customer\_data['PCA1'] = pca\_components[:, 0]*  *customer\_data['PCA2'] = pca\_components[:, 1]*  *plt.figure(figsize=(10, 6))*  *sns.scatterplot(x='PCA1', y='PCA2', hue='Cluster', palette='Set2', data=customer\_data, s=100, alpha=0.7)*  *plt.title('Customer Segmentation (Clusters) - PCA')*  *plt.xlabel('PCA1')*  *plt.ylabel('PCA2')*  *plt.legend()*  *plt.show()* |

Clustering Report:

* Number of clusters formed.
* DB Index and Silhouette Score values.