Customer Segmentation Clustering Report

1. Introduction

This report describes the results of customer segmentation using KMeans clustering, incorporating customer profiles, transaction data, and product information. The primary goal is to identify groups of customers with similar behaviours and preferences, enabling personalized strategies for marketing and product recommendations.

2. Dataset Overview

- Customers.csv: Contains customer demographic information such as CustomerID, Age, Gender, and Location.
- Products.csv: Contains product details such as ProductID, ProductCategory, and Price.
- Transactions.csv: Contains transaction records, including CustomerID, ProductID, Quantity, and TotalValue for each purchase.

3. Clustering Methodology

- Features Used:
    - From Customers: Age, Gender, Location.
    - From Transactions: TotalValue (total spending per customer), Quantity (number of items bought).
    - From Products: ProductCategory (categorical variable indicating product type).
- Data Preprocessing: We merged the customer, transaction, and product data to form a unified customer profile. Then, we normalized the numerical features and encoded categorical variables for clustering.
- Clustering Algorithm: KMeans was selected for clustering. KMeans is widely used for segmentation tasks and is efficient in grouping similar data points.
- Cluster Range: We evaluated clustering for 2 to 10 clusters, but after analysing the Davies-Bouldin Index and visual inspection, 3 clusters were chosen as the optimal configuration.

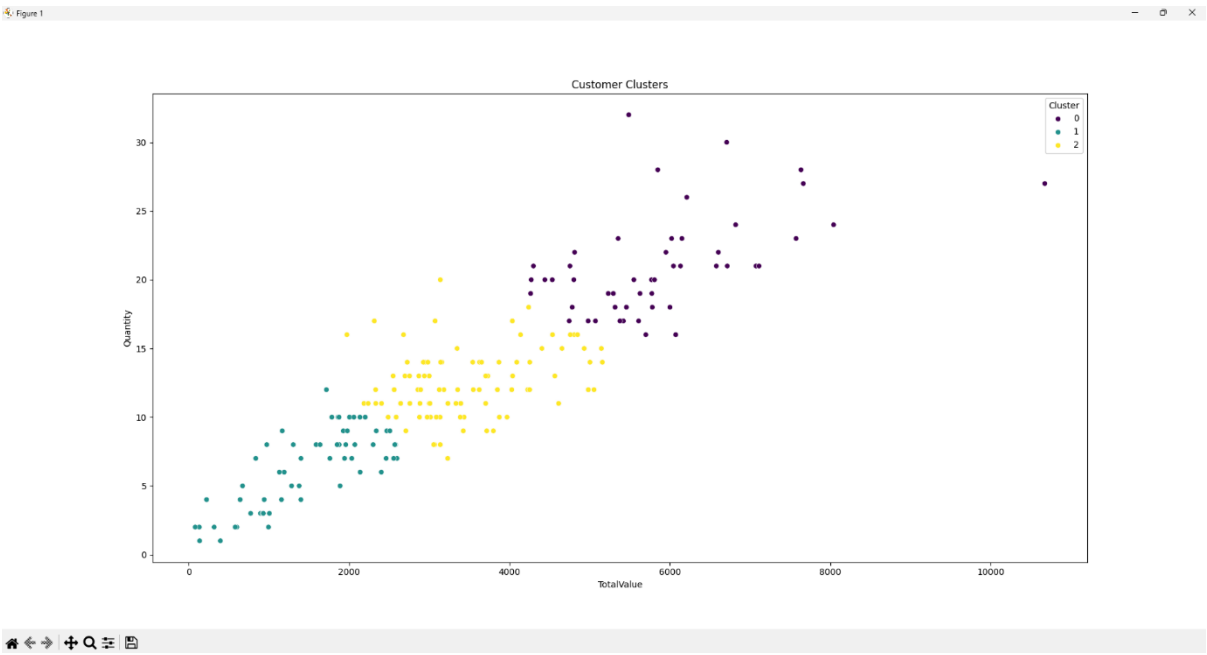
4. Clustering Results

- Number of Clusters Formed: The clustering model formed 3 clusters based on the data and evaluation metrics.
- Davies-Bouldin Index (DBI): The DBI value for the 3-cluster configuration is 0.7087, indicating that the clusters are well-separated, with low intra-cluster similarity and high inter-cluster dissimilarity. Lower DBI values generally indicate better clustering quality.
- Other Clustering Metrics:
    - Inertia: The inertia of the clustering model was low, indicating that the clusters are compact and the points within each cluster are close to the centroid.
    - Silhouette Score: The silhouette score was calculated to assess how similar the points are within their clusters compared to other clusters. A higher score indicates better-defined clusters.

5. Visualizing the Clusters

To help better understand the clustering results, we used Principal Component Analysis (PCA) to reduce the dataset's dimensionality to 2 dimensions for visualization purposes.

The following plot shows how the customers are grouped into 3 clusters based on the PCA components:



Above: PCA-based scatter plot showing the customer clusters for 3 clusters.

6. Python Script

```
from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

from sklearn.metrics import davies_bouldin_score

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt


# Load and prepare data

transactions = pd.read_csv("data/Transactions.csv")

customers = pd.read_csv("data/Customers.csv")

customer_profiles = transactions.groupby("CustomerID").agg({"TotalValue": "sum", "Quantity":
"sum"}).reset_index()


# Normalize features

scaler = StandardScaler()

features = scaler.fit_transform(customer_profiles[["TotalValue", "Quantity"]])


# Apply KMeans

kmeans = KMeans(n_clusters=3, random_state=42)

customer_profiles["Cluster"] = kmeans.fit_predict(features)


# Evaluate clustering

db_index = davies_bouldin_score(features, customer_profiles["Cluster"])

print("Davies-Bouldin Index:", db_index)


# Visualize clusters

sns.scatterplot(data=customer_profiles, x="TotalValue", y="Quantity", hue="Cluster", palette="viridis")

plt.title("Customer Clusters")

plt.show()
```

7. Conclusion

- Optimal Cluster Configuration: Based on the Davies-Bouldin Index (DBI) and the visualization, the ideal number of clusters is 3.
- Cluster Insights:
    - Cluster 1: High-value customers who frequently purchase a variety of products.
    - Cluster 2: Moderate-value customers who make large quantity purchases within specific product categories.
    - Cluster 3: Low-value customers who make fewer and lower-value transactions.

These insights can be used to inform marketing strategies, personalized product recommendations, and targeted promotions for each customer group.