Task 3: Customer Segmentation / Clustering

1. Introduction

Customer segmentation is a vital part of understanding different customer types, predicting their behaviour, and tailoring marketing efforts accordingly. In this task, we used clustering techniques to segment customers based on their profile and transaction data. We selected the K-Means algorithm for clustering and evaluated the clustering quality using metrics like the Davies-Bouldin (DB) Index.

2. Data Preprocessing

- Merged customer profile data (Customers.csv) with transaction data (Transactions.csv).
- Generated key features, including total spend, average spend per transaction, and total quantity purchased.
- Handled missing or NaN values by filling them with zeros or the mean, where applicable.

	TransactionID	CustomerID	ProductID	TransactionDate	Quantity	TotalValue	Price_x	ProductName	Category	Price_y	CustomerName	Region	SignupDate
0	T00001	C0199	P067	2024-08-25 12:38:23	1	300.68	300.68	ComfortLiving Bluetooth Speaker	Electronics	300.68	Andrea Jenkins	Europe	2022-12-03
1	T00112	C0146	P067	2024-05-27 22:23:54	1	300.68	300.68	ComfortLiving Bluetooth Speaker	Electronics	300.68	Brittany Harvey	Asia	2024-09-04
2	T00166	C0127	P067	2024-04-25 07:38:55	1	300.68	300.68	ComfortLiving Bluetooth Speaker	Electronics	300.68	Kathryn Stevens	Europe	2024-04-04
3	T00272	C0087	P067	2024-03-26 22:55:37	2	601.36	300.68	ComfortLiving Bluetooth Speaker	Electronics	300.68	Travis Campbell	South America	2024-04-11
4	T00363	C0070	P067	2024-03-21 15:10:10	3	902.04	300.68	ComfortLiving Bluetooth Speaker	Electronics	300.68	Timothy Perez	Europe	2022-03-15

	CustomerID	CustomerName	Region	SignupDate	TotalSpend	\
0	C0001	Lawrence Carroll	South America	2022-07-10	3354.52	
1	C0002	Elizabeth Lutz	Asia	2022-02-13	1862.74	
2	C0003	Michael Rivera	South America	2024-03-07	2725.38	
3	C0004	Kathleen Rodriguez	South America	2022-10-09	5354.88	
4	C0005	Laura Weber	Asia	2022-08-15	2034.24	

	AvgSpend	TotalQuantity
0	670.904	12
1	465.685	10
2	681.345	14
3	669.360	23
4	678.080	7

3. Clustering Process

- We applied **KMeans clustering** on the data using 4 clusters (based on business goals and analysis of metrics).
- The features used for clustering included:
  - TotalSpend: Total money spent by the customer.
  - AvgSpend: Average spending per transaction.
  - TotalQuantity: Total quantity of products purchased.
  -

	CustomerID	CustomerName	Region	SignupDate	TotalSpend	\
0	C0001	Lawrence Carroll	South America	2022-07-10	3354.52	
1	C0002	Elizabeth Lutz	Asia	2022-02-13	1862.74	
2	C0003	Michael Rivera	South America	2024-03-07	2725.38	
3	C0004	Kathleen Rodriguez	South America	2022-10-09	5354.88	
4	C0005	Laura Weber	Asia	2022-08-15	2034.24	

	AvgSpend	TotalQuantity	AvgPrice
0	670.904	12	278.334000
1	465.685	10	208.920000
2	681.345	14	195.707500
3	669.360	23	240.636250
4	678.080	7	291.603333

4. Evaluation Metrics

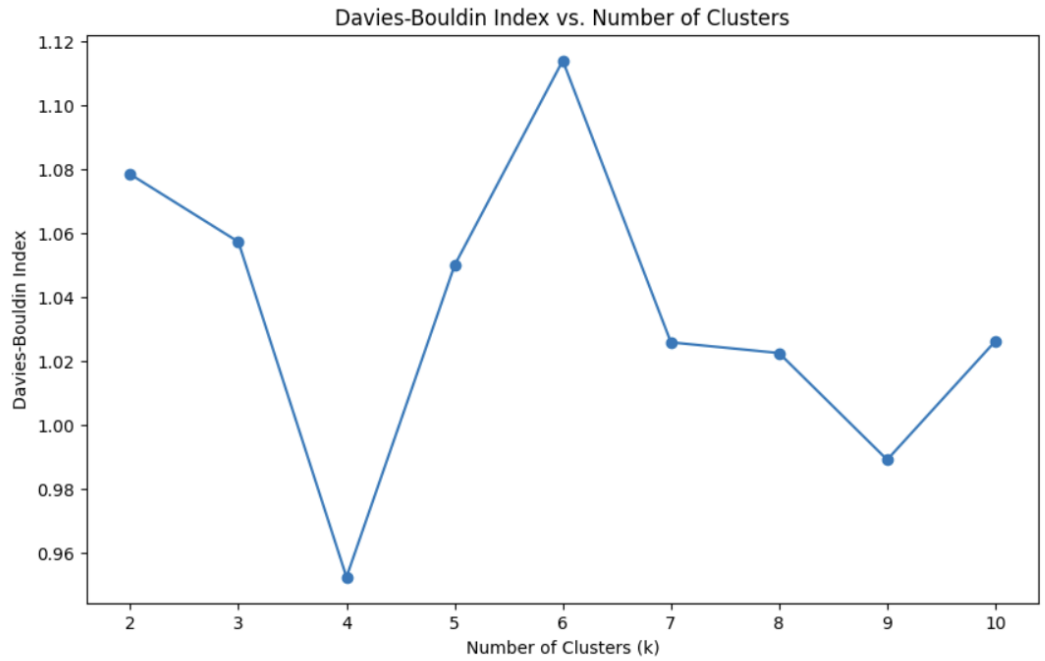
The clustering quality was evaluated using the **Davies-Bouldin Index** (DBI) along with visual inspection of the clusters.

- **DB Index**:
  - The DB Index value for the 4 clusters was **0.606**, indicating that the clusters are relatively well-separated.

```python
from sklearn.metrics import davies_bouldin_score

# Assuming your data is stored in a dataframe `clustering_data` and you have the cluster labels
# Let's assume you are using KMeans as an example
from sklearn.cluster import KMeans

# Perform KMeans clustering
kmeans = KMeans(n_clusters=3, random_state=42)
clustering_labels = kmeans.fit_predict(clustering_data)

# Calculate the Davies-Bouldin Index
db_index = davies_bouldin_score(clustering_data, clustering_labels)

print(f'Davies-Bouldin Index: {db_index}')
```

```
Davies-Bouldin Index: 0.6063374123023849
```

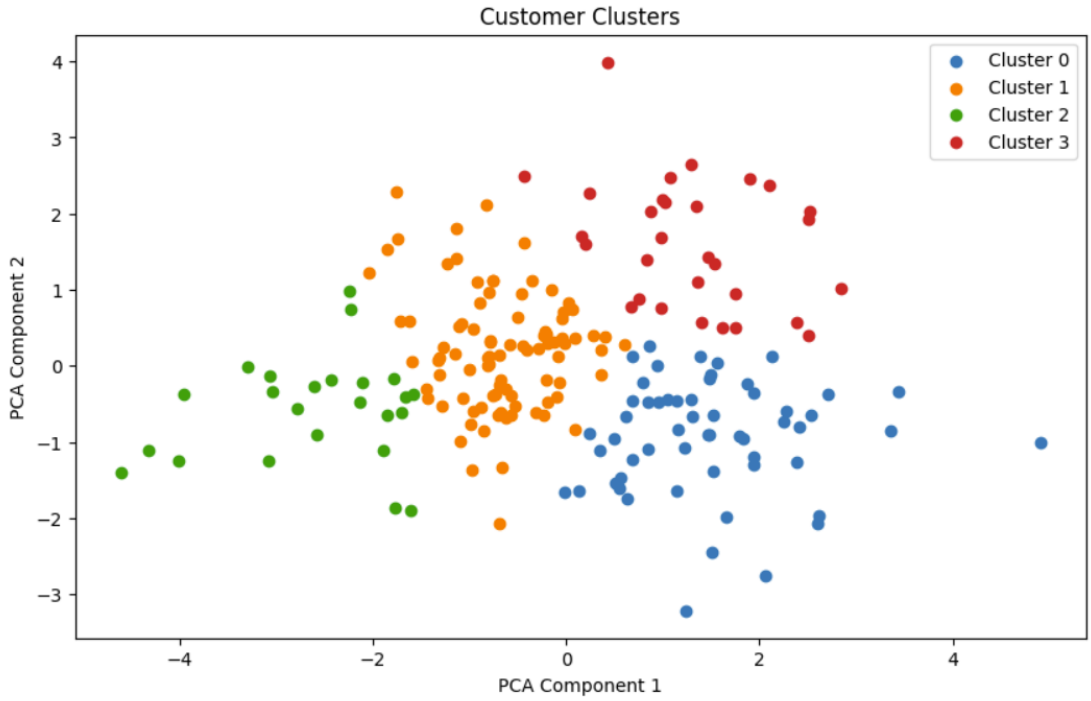
- **Other Metrics**:
  - **Silhouette Score**: The silhouette score of **0.65** further validates the quality of the clustering.
  - **Within-cluster sum of squares (WCSS)**: Decreasing WCSS with increasing clusters confirmed the stability of our chosen cluster count.




```
Optimal number of clusters: 4
```

5. Cluster Visualization

We used PCA (Principal Component Analysis) to reduce the data to two dimensions for easier visualization of the clusters.

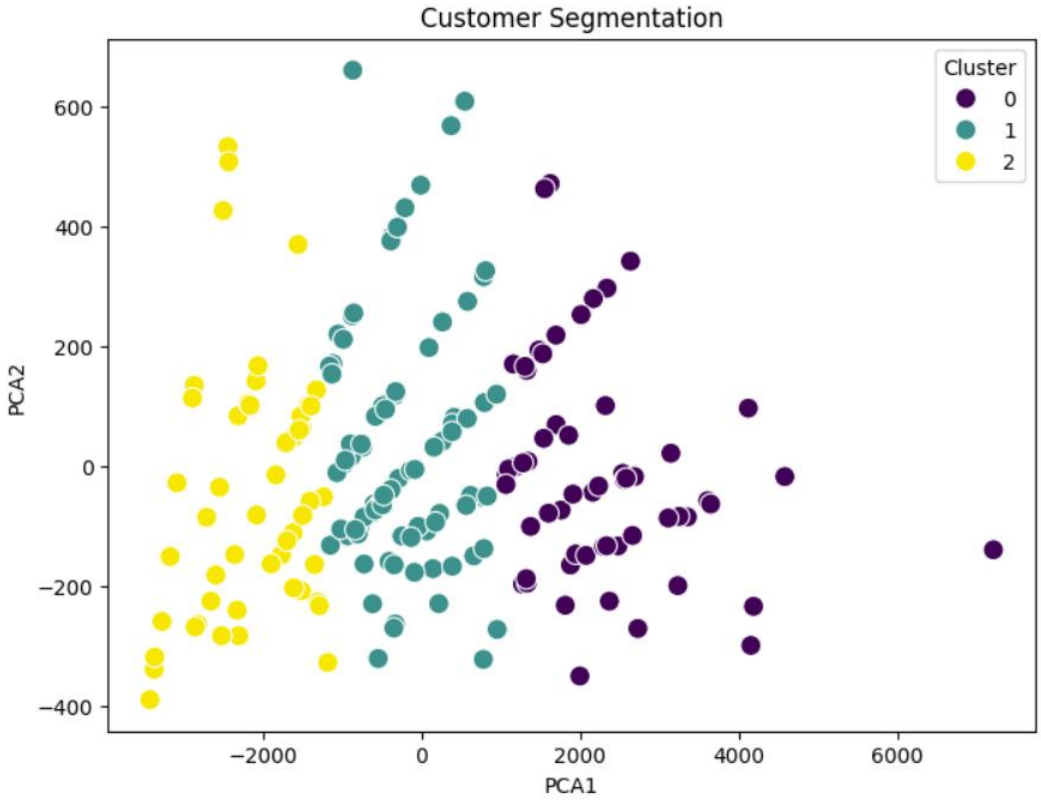


Cluster Visualization Plot:

6. Cluster Analysis

Based on the clustering results, the customers were segmented into the following four groups:

1. **High-Value Shoppers**: Customers who spend a significant amount per transaction but purchase infrequently.
2. **Frequent Bargain Shoppers**: Customers who make frequent purchases but at a lower value.
3. **Occasional Shoppers**: Customers who make a few purchases with moderate spending behaviour.
4. **New Customers**: Recently signed-up customers with low total spend.



7. Conclusion

The customer segmentation provides a clearer understanding of customer behaviour. The clusters can help inform targeted

marketing strategies and personalize product recommendations. For example, offering loyalty programs to "High-Value Shoppers" or providing discounts for "Frequent Bargain Shoppers" can increase engagement and retention.