Task 3: Customer Segmentation Report

1. Introduction

- **Objective**: The aim of this analysis is to segment customers using clustering techniques by leveraging profile information from Customers.csv and transaction data from Transactions.csv.
- **Methods Used**: Two clustering algorithms, *KMeans* and *Hierarchical Clustering*, were applied. Clustering metrics such as *Davies-Bouldin Index (DBI)* and *Silhouette Score* were used to evaluate the results.
- **Tools & Libraries**: Python, pandas, scikit-learn, seaborn, matplotlib, and scipy.

---

2. Data Preparation

- **Data Sources**:
  - *Customers.csv*: Contains customer profile information (e.g., CustomerID, Region, Signup Date, etc.).
  - *Transactions.csv*: Includes transactional data (e.g., Total Value, Quantity)
  -
- **Feature Engineering**:
  - Merged both datasets using the CustomerID column.
  - Aggregated transactional data to compute features such as:
    - total_spending: Total value of all transactions per customer.
    - avg_order_value: Average transaction value per customer.
    - transaction_count: Number of transactions per customer.
    - total_quantity: Total quantity purchased per customer.
  - One-hot encoded categorical features like Region.
  -

- **Normalization**:
  - Applied **StandardScaler** to normalize numerical features to a mean of 0 and a standard deviation of 1.

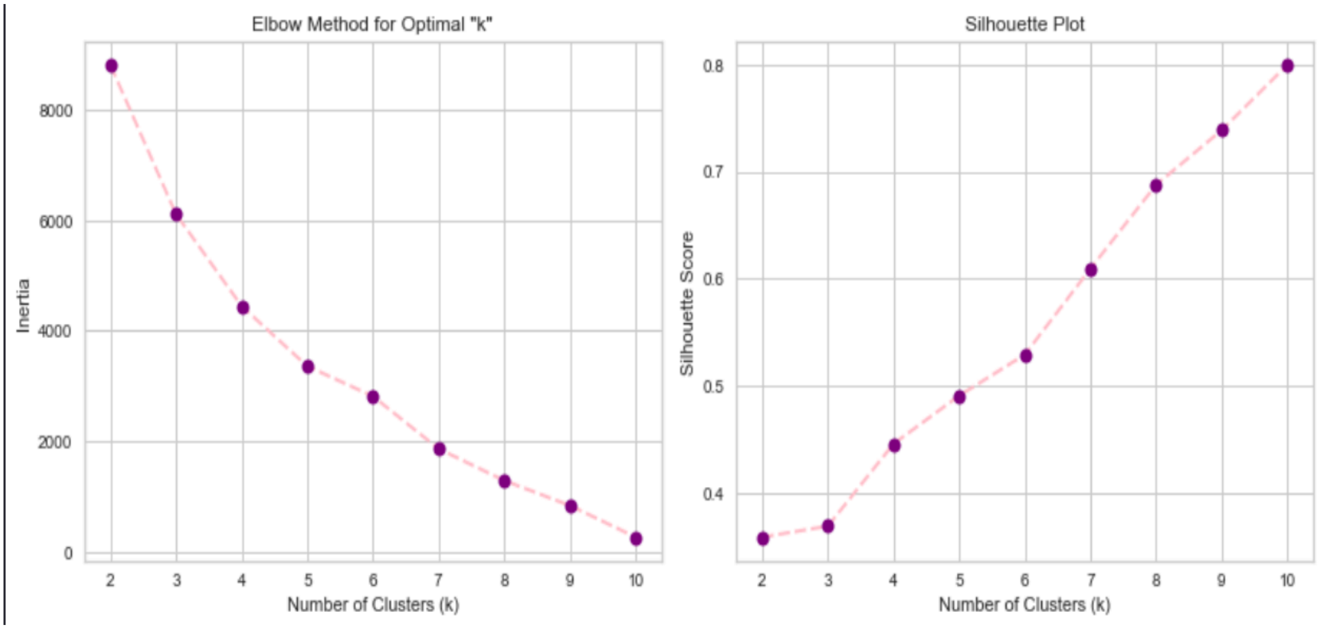
---

3. Clustering Analysis

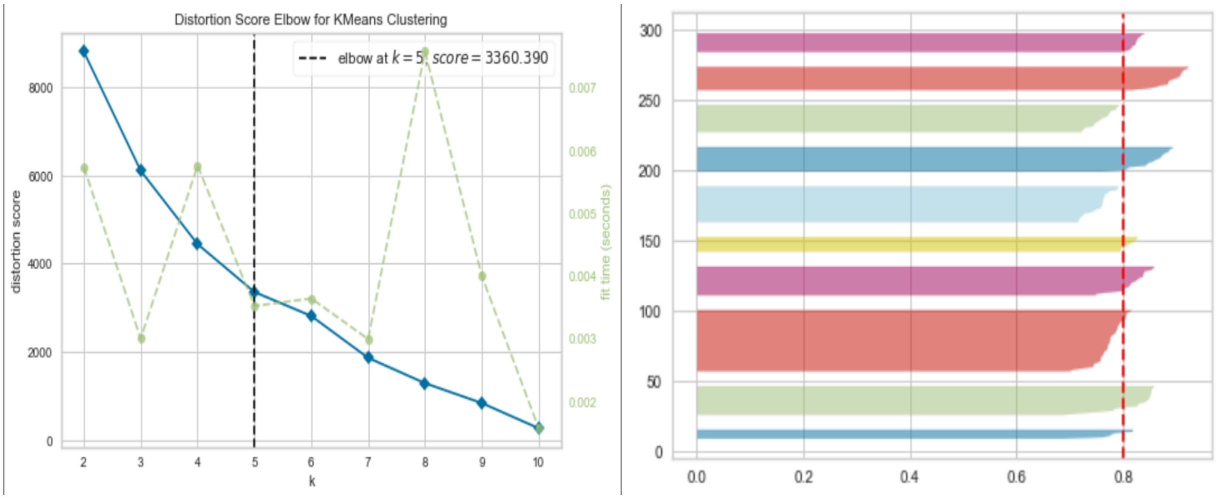
3.1 KMeans Clustering

- *Optimal Number of Clusters*: Using the Elbow Method and Silhouette Scores, the optimal number of clusters was determined to be **10**.
- *Metrics:*
  - Silhouette Score: **0.7994**
  - Davies-Bouldin Index (DBI): **0.2791**
- *Visualizations:*

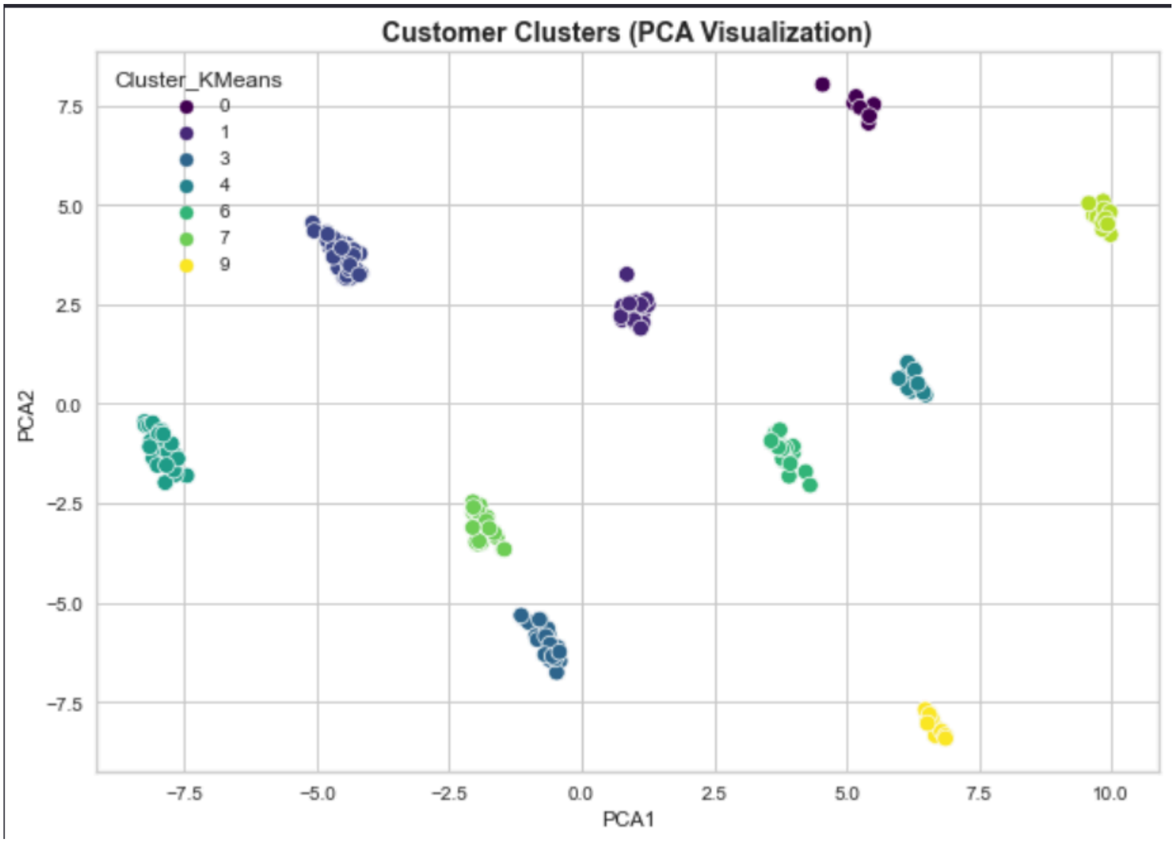
Elbow Curve & Silhouette Plot:



Yellowbrick Elbow Curve & Silhouette Plot:



Cluster Visualization: A PCA-based 2D scatterplot showing the clusters is provided below:

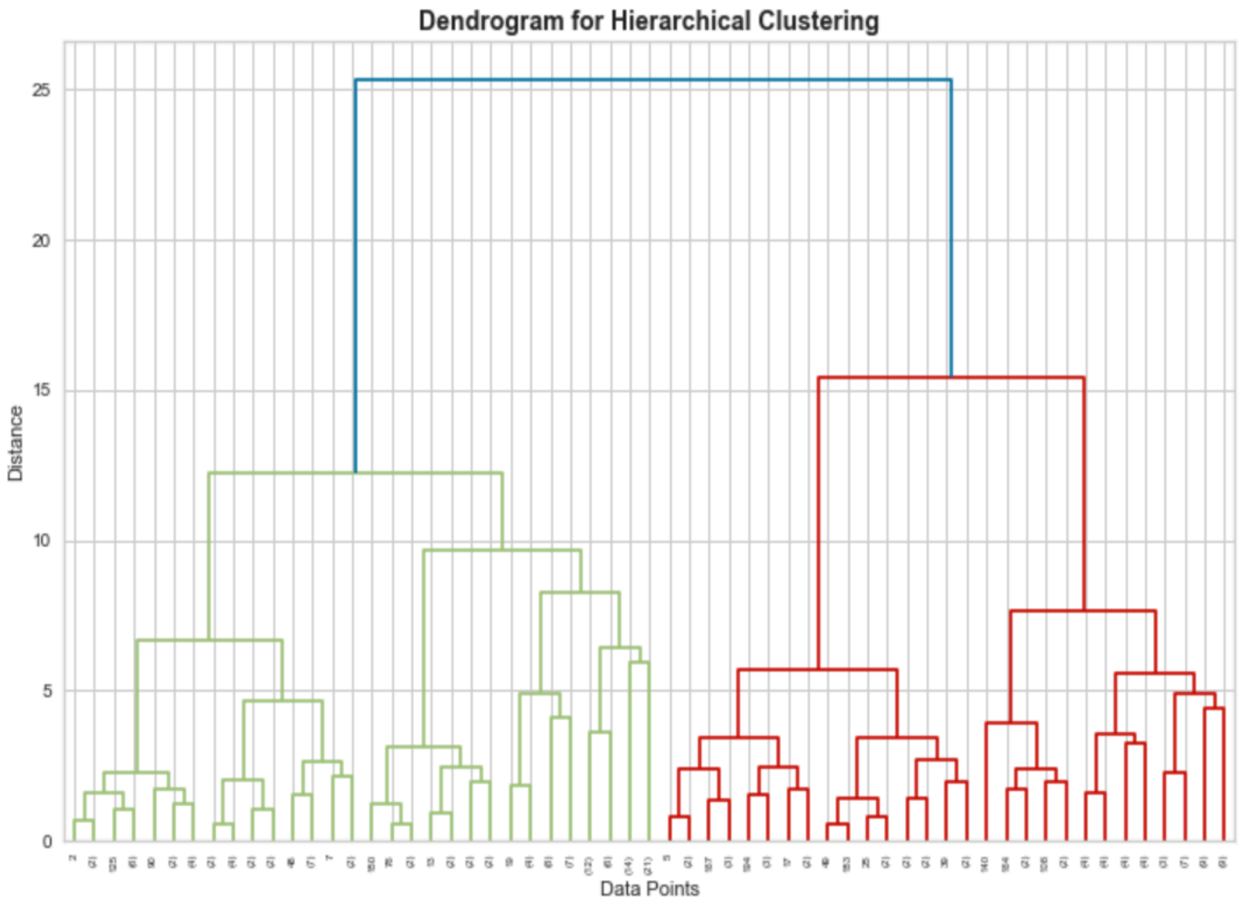


3.2 Hierarchical Clustering

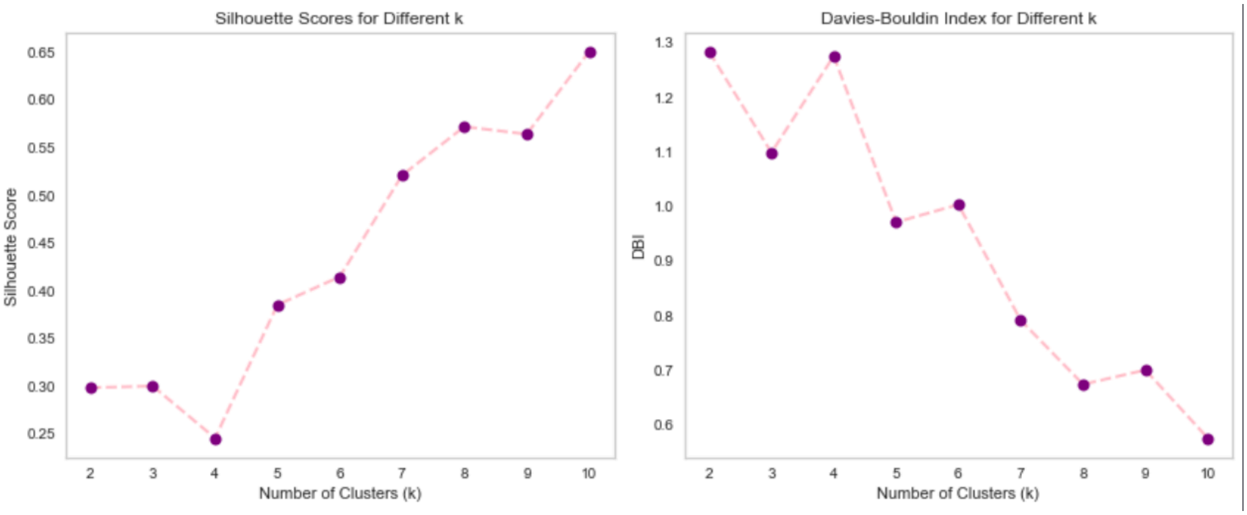
- *Optimal Number of Clusters:* Evaluated using Dendrogram and metrics; the optimal number of clusters was **10**.

- *Metrics:*
  - Silhouette Score: **0.7419**
  - Davies-Bouldin Index (DBI): **0.3675**
- *Visualizations:*

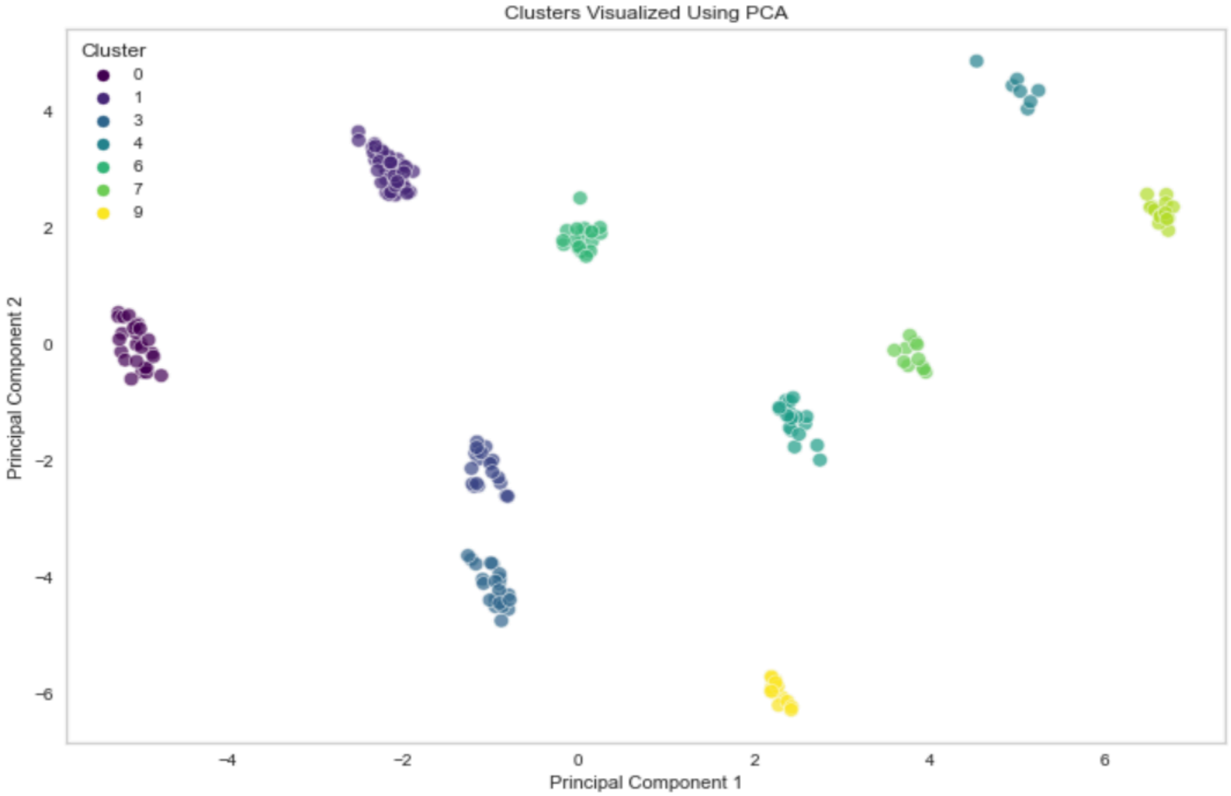
Dendrogram:



Silhouette Score and DBI plot:



Cluster Visualization: A PCA-based 2D scatterplot showing the clusters is provided below



4. Comparison of Results

A comparison of clustering metrics between KMeans and Hierarchical Clustering is shown below:

Algorithm	**Silhouette Score**	**DBI (Davies-Bouldin Index)**
KMeans	0.799487	0.279196
Hierarchical	0.741938	0.367553

---

5. Conclusion

Key Findings:

- KMeans with 10 clusters resulted in a better Davies-Bouldin Index (0.279).
- Hierarchical Clustering achieved comparable performance but with slightly lower metrics.

Recommendations: KMeans is recommended for customer segmentation in this case due to its better clustering metrics.

---

6. Appendix

Github Repo Link:
https://github.com/faizanxmulla/zeotap-data-science-intern-assignment