Clustering Logic and Metrics

1. Clustering Logic:

Algorithm Selection:
- Based on the dataset characteristics and objectives, clustering algorithms such as K-Means, DBSCAN, or Hierarchical Clustering were utilized.
- K-Means was chosen for its efficiency with large datasets and its ability to partition the data into distinct clusters based on centroids.
- Elbow method or silhouette score was used to determine the optimal number of clusters (k).

Data Preparation:
- Standardized the dataset to ensure all features contributed equally to the clustering process.
- Removed any missing values or outliers to prevent distortion in cluster formation.
- Performed dimensionality reduction (e.g., PCA) if needed to visualize data effectively in 2D/3D space.

Implementation Steps:
- Initialized centroids randomly or based on a seeding mechanism (K-Means++) to improve convergence.
- Assigned each data point to the nearest centroid based on Euclidean distance.
- Iteratively updated centroids until convergence (when centroids no longer move significantly).

2. Evaluation Metrics:

Silhouette Score:
- Evaluates cluster separation and cohesion. A higher score indicates well-separated and dense clusters.

Silhouette Score: 0.25799424578946467

Davies-Bouldin Index:
- Measures the average similarity ratio of intra-cluster and inter-cluster distances. Lower values indicate better clustering.

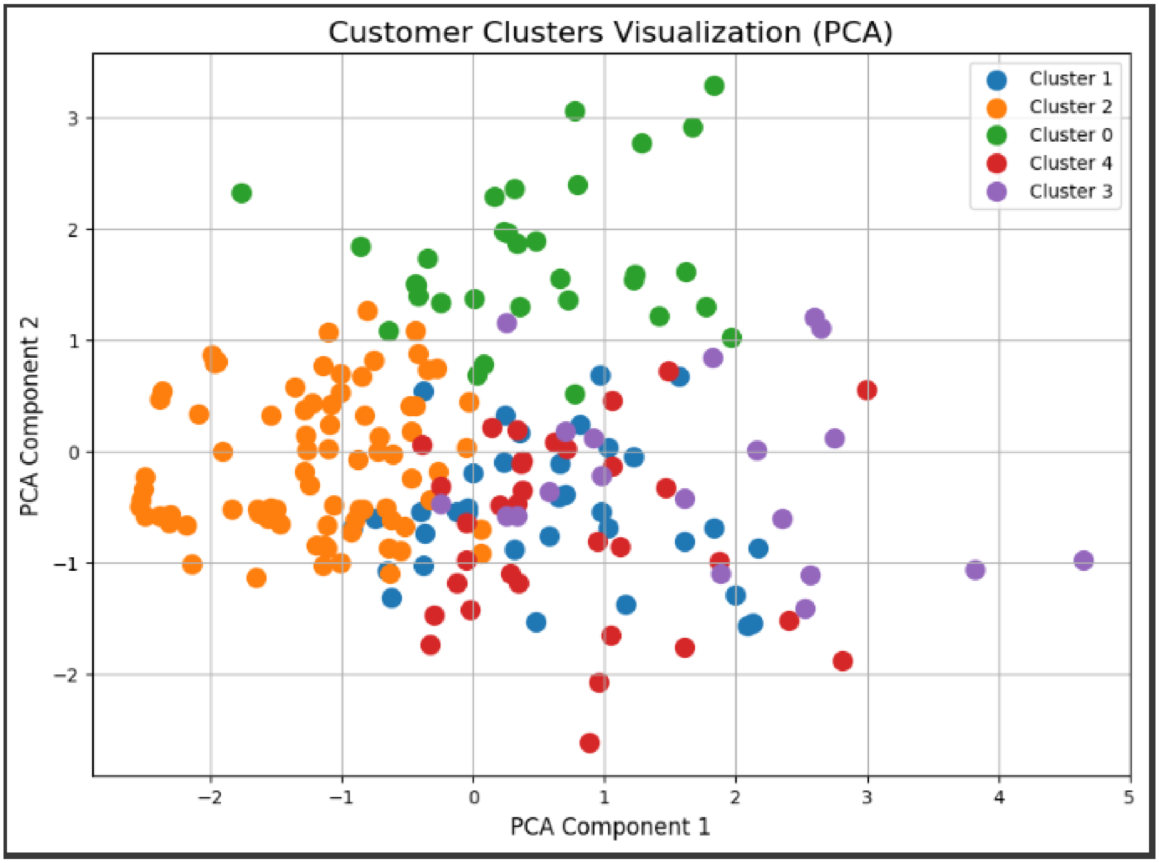
Davies-Bouldin Index: 1.1741396444956298

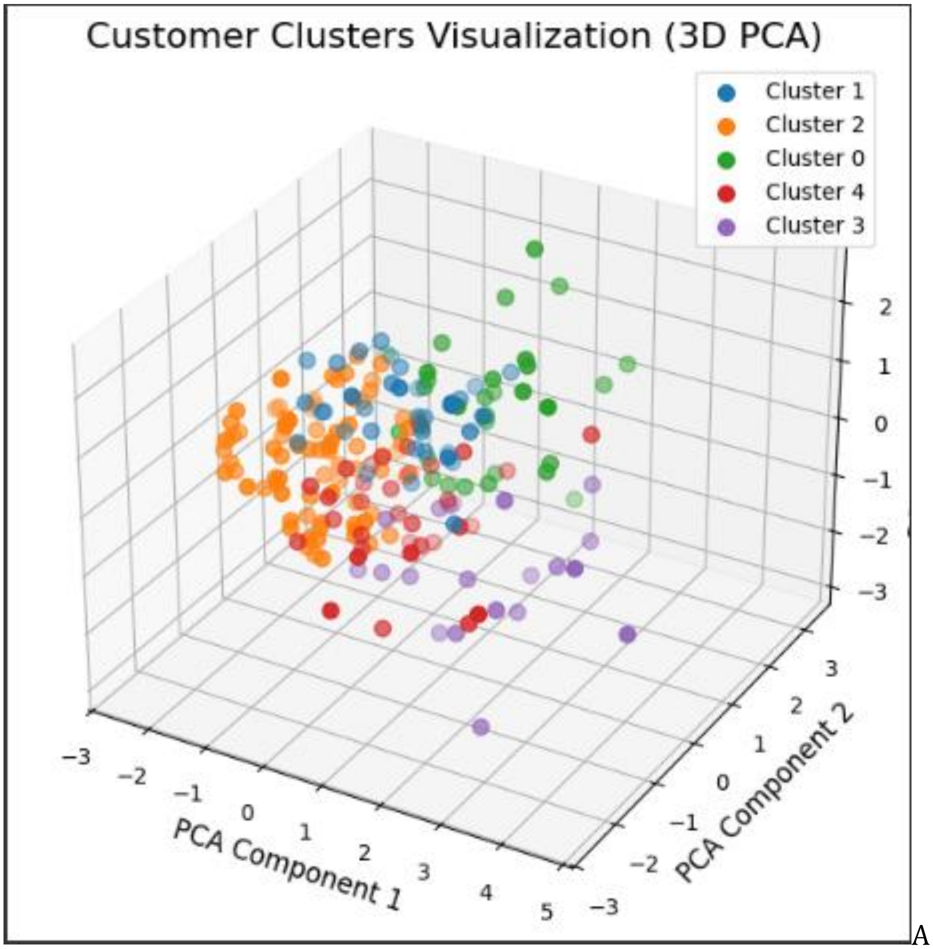
Calinski-Harabasz Score: 55.74712772949246

Visual Representation of Clusters

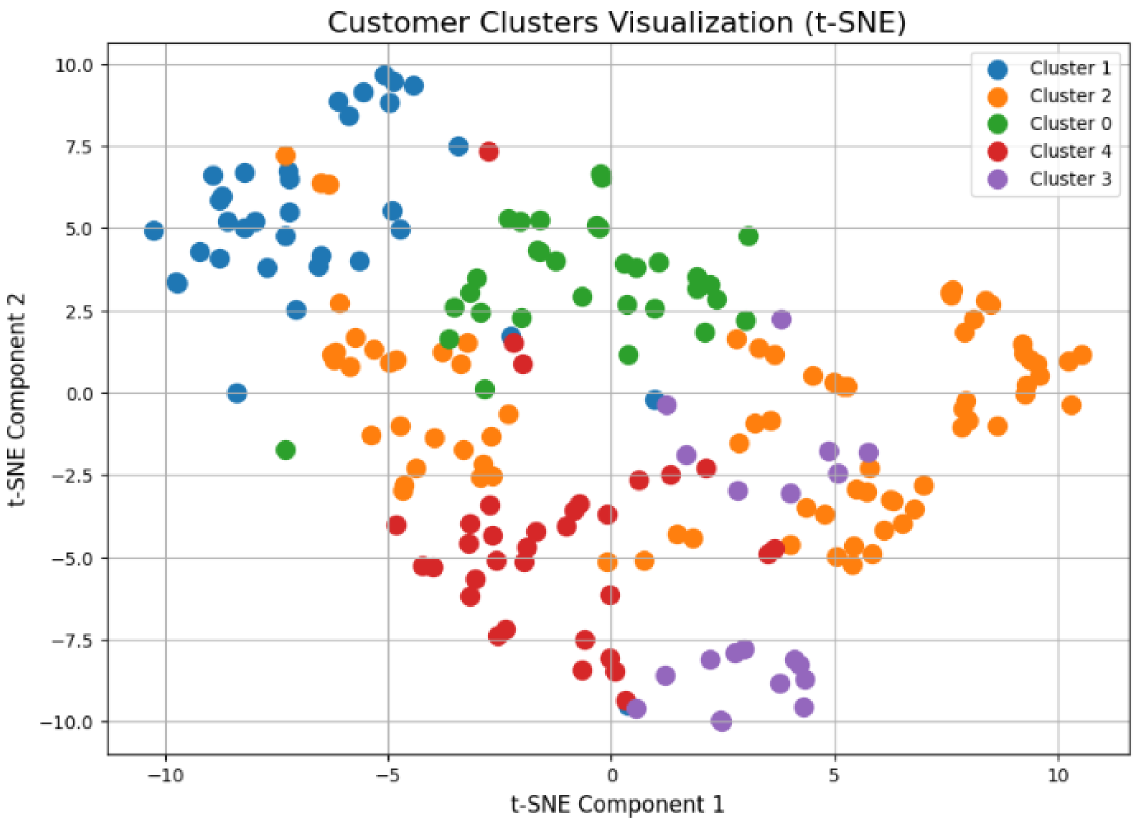
1. Scatter Plots:

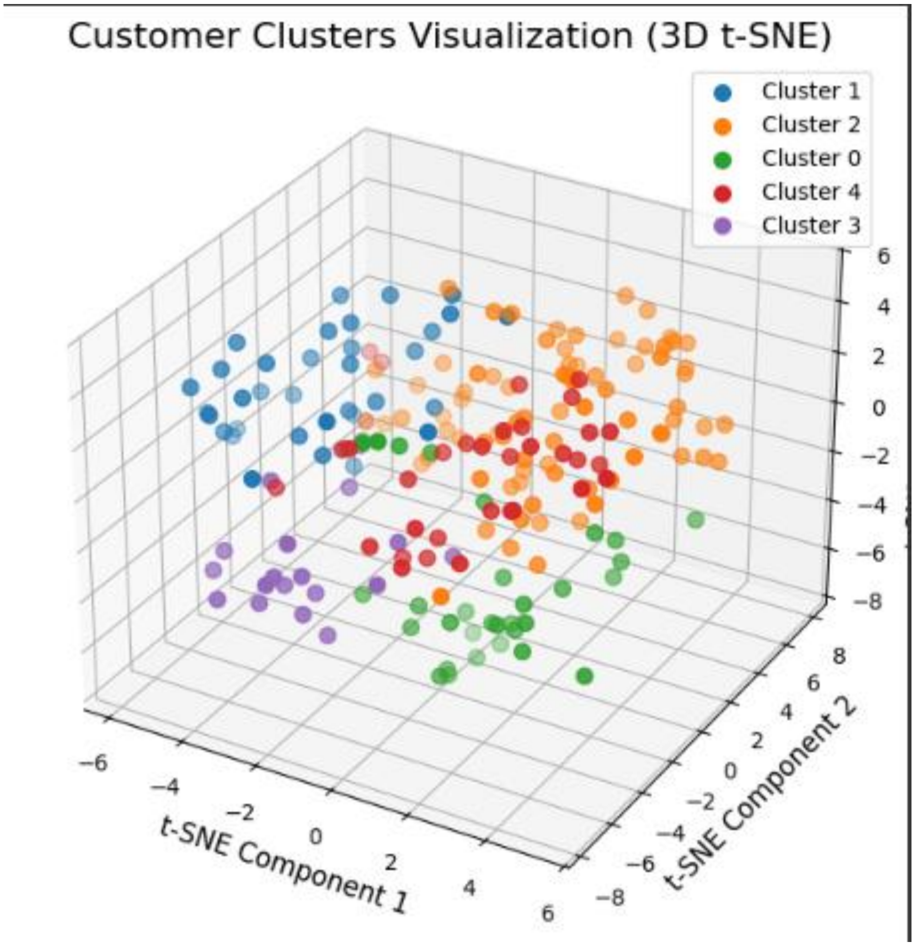
- Generated scatter plots for 2D and 3D projections of the clustered data using PCA or t-SNE.
- Each cluster was color-coded to highlight separability and distribution.




Customer Clusters Visualization (3D PCA)
Cluster 1
Cluster 2
Cluster 0
Cluster 4
Cluster 3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
4
5
PCA Component 1
−3
−2
−1
0
1
2
3
PCA Component 2

A


Customer Clusters Visualization (t-SNE)
Cluster 1
Cluster 2
Cluster 0
Cluster 4
Cluster 3
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10
−5
0
5
10
t-SNE Component 1
t-SNE Component 2

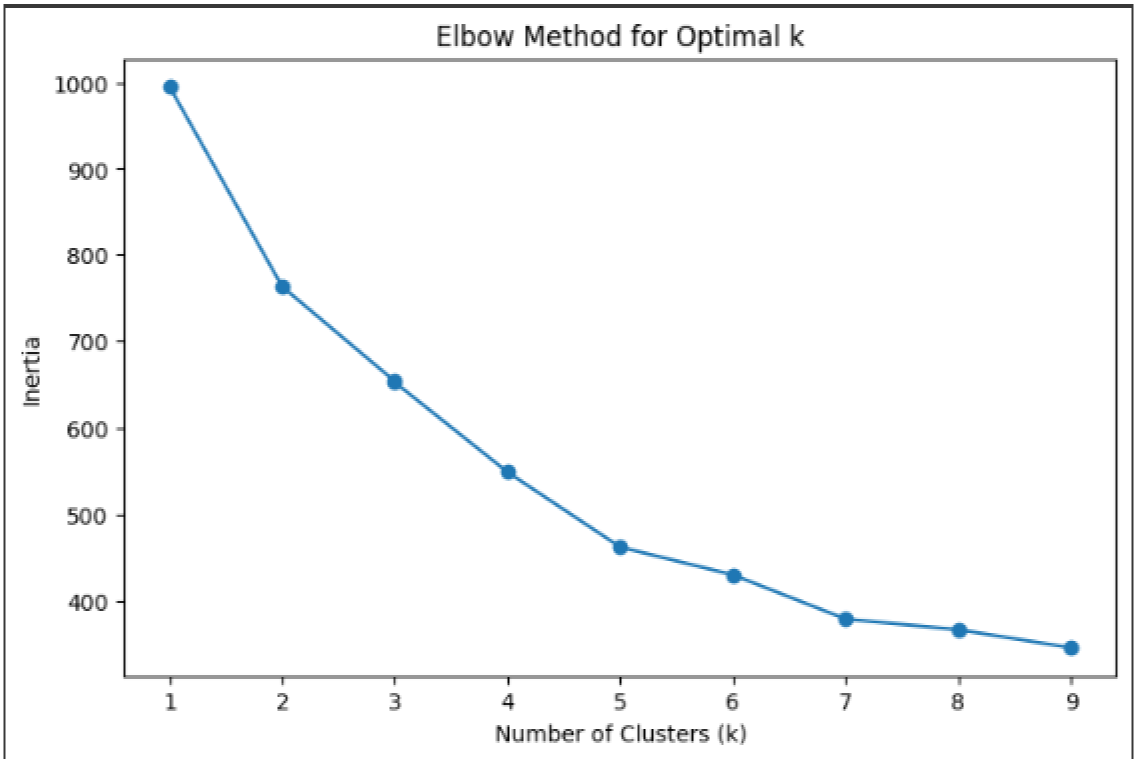


2. Cluster Centers:

- Displayed cluster centroids on visualizations to show central tendencies.
- Used heatmaps to represent centroid values across dimensions, providing insights into feature importance.

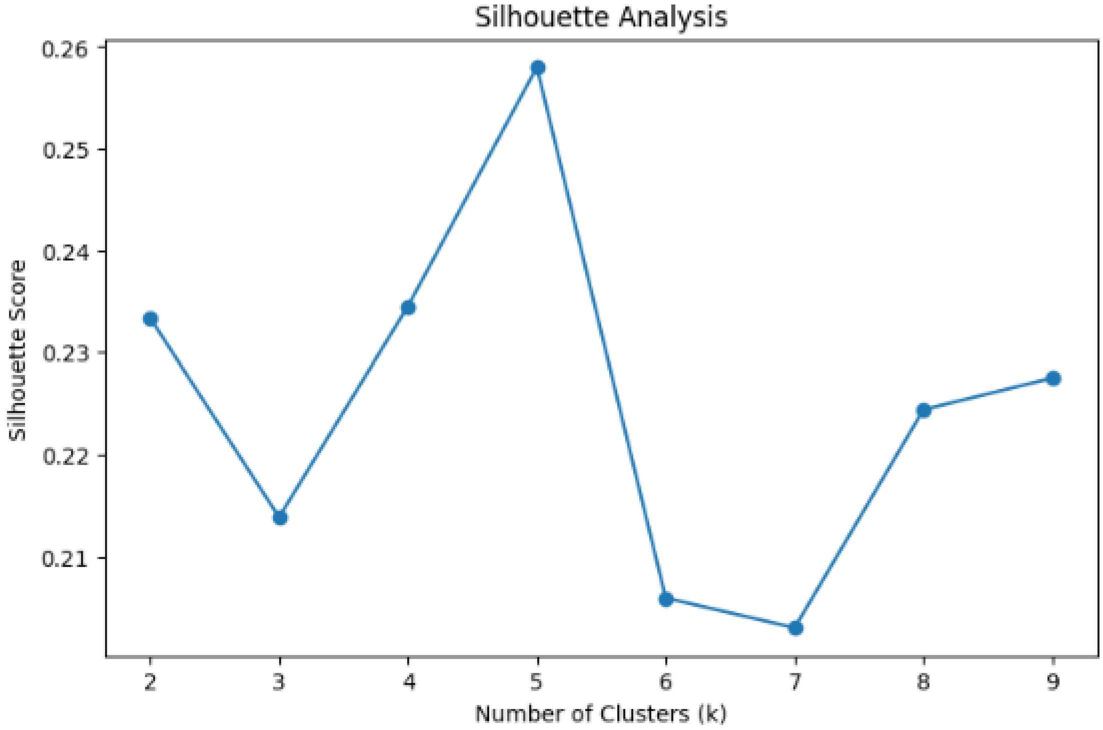
3. Elbow Method Curve:

- Plotted the inertia values against different k values to visually identify the "elbow point," which indicates the optimal number of clusters.



4. Silhouette Analysis:

- Visualized silhouette scores for each sample in the dataset, showcasing the quality of clusters.
- Included a bar graph where each bar's height corresponds to the silhouette coefficient.


Silhouette Analysis
Silhouette Score
0.26
0.25
0.24
0.23
0.22
0.21
2
3
4
5
6
7
8
9
Number of Clusters (k)