Customer Segmentation Using K-Means

Retail/E-commerce Case Study

By Tanvir Bakther

Objective

- The purpose of this project is to analyze customer behavior using unsupervised learning methods and segment them into meaningful groups. The goal is to help retail or e-commerce businesses create targeted marketing campaigns, improve customer experience, and optimize product offerings based on cluster-specific characteristics.
- We used K-means clustering, a popular unsupervised machine learning algorithm, to categorize customers based on demographic and behavioral data.

Dataset

We utilized a sample dataset containing 10 customer records, each with the following features:

CustomerID – A unique identifier for each customer.

Gender – Categorical feature with values 'Male' or 'Female'.

Age – Customer's age in years.

Annual Income (k$) – Customer's annual income in thousands of dollars.

Spending Score (1–100) – A value assigned by the store to each customer based on spending patterns and behavior.

Data Preprocessing

To prepare the data for clustering:

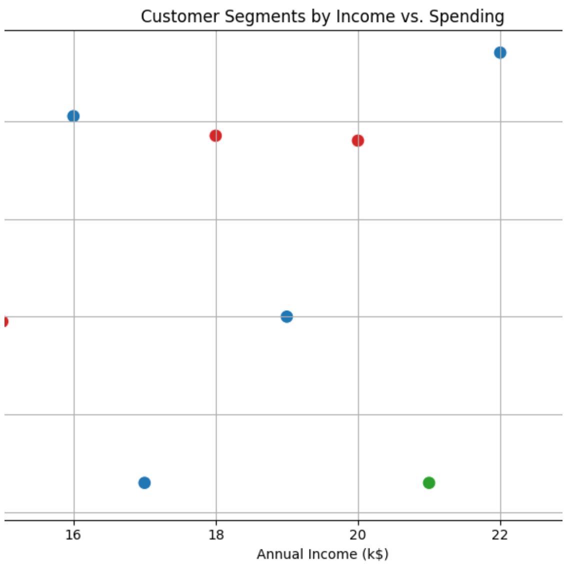
- **Label Encoding**: We encoded the 'Gender' column into numeric values using 0 for 'Female' and 1 for 'Male'.
- **Feature Selection**: We selected 'Gender', 'Age', 'Annual Income (k$)', and 'Spending Score (1–100)' for clustering.
- **Normalization**: The selected features were scaled using **StandardScaler** from scikit-learn to ensure all variables contribute equally to the clustering process.

K-Means Clustering

Algorithm Steps:

- Randomly initialize k centroids.
- Assign each customer to the nearest centroid using Euclidean distance.
- Compute new centroids by averaging the points in each cluster.
- Repeat until centroids no longer change significantly (convergence).
- We experimented with different values of **k** and evaluated the models using:
- Inertia (within-cluster sum of squares)
- Silhouette Score (cluster separation and cohesion)

Annual Income vs. Spending Score



Scatter plots of customers by Annual Income vs. Spending Score, colored by cluster.

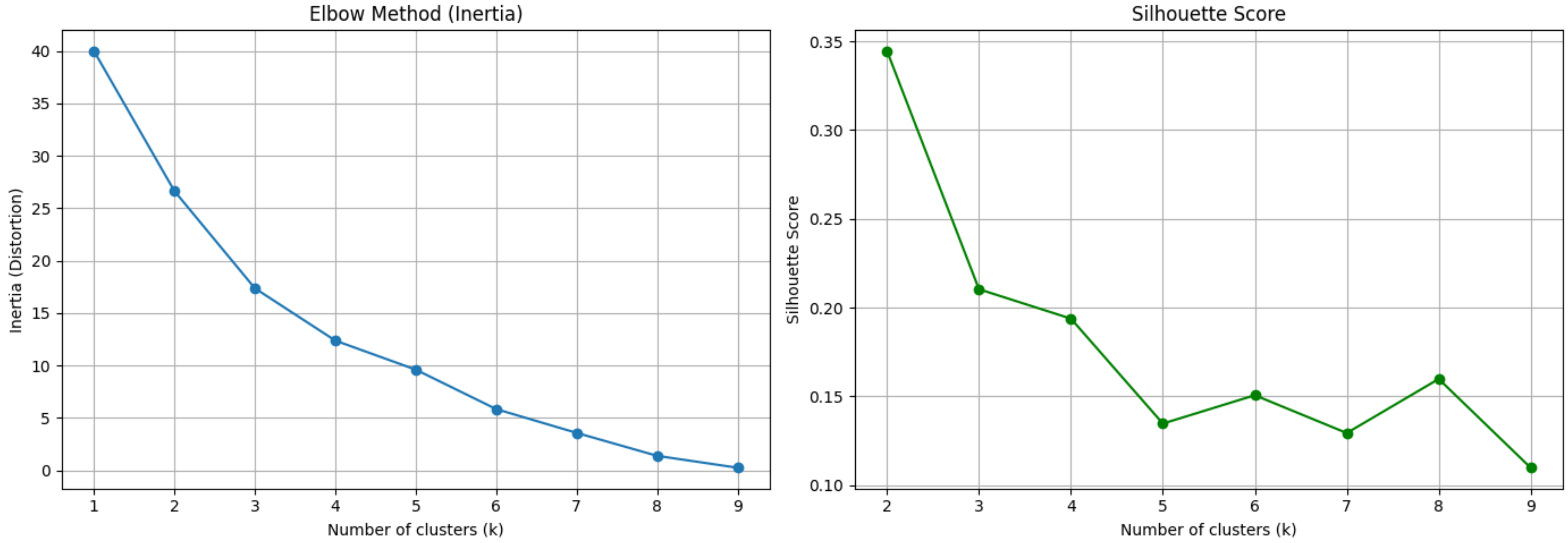
Elbow and Silhouette Method Analysis

We tried values of k = 1 to 9:

- Elbow Method showed the curve bending at k = 3 and k = 4, indicating good potential values.
- Silhouette Score peaked at k = 3, meaning the clusters are well-separated and dense.

Cluster Analysis (k = 3)

- We experimented with different values of **k** and evaluated the models using:
- **Inertia** (within-cluster sum of squares)
- **Silhouette Score** (cluster separation and cohesion)



Result and Interpretation

- **Clustering with k = 4:**
- **Cluster 0**: Young males with mid-low income and **high spending** — possible brand-loyal or trend-driven shoppers.
- **Cluster 1**: Older female customer with high income but **very low spending** — possibly budget-conscious or inactive.
- **Cluster 2**: Young females with low income but **moderate to high spending** — may represent aspirational or loyal shoppers.
- **Cluster 3**: Middle-aged males with decent income and **low spending** — occasional or non-impulsive buyers.
- **Clustering with k = 3:**
- **Cluster A**: High spenders with varied income — good candidates for VIP programs.
- **Cluster B**: Low spenders — might need promotional incentives.
- **Cluster C**: Moderate income and moderate spending — stable, average customers.

Conclusion

- The K-means clustering approach successfully grouped customers into clear segments. Both **k = 3 and k = 4** proved useful:
- **k = 3** is more interpretable and simpler for business action.
- **k = 4** provides more nuanced segmentation.

Tools Used

- Python (Pandas, NumPy)
- scikit-learn (KMeans, StandardScaler, Silhouette Score)
- matplotlib & seaborn (for visualizations)