Project title: Customer Segmentation using Clustering Algorithms

Submitted By

Name	Student ID
Md.Mahabub Rana	0242220005101234

This Report is presented in Partial Fulfillment of the course CSE 325: Data Mining and Machine Learning in the Computer Science and Engineering Department

DAFFODIL INTERNATIONAL UNIVERSITY

Dhaka, Bangladesh

Submission date: August 11, 2025

DECLARATION

We hereby declare that we have done this lab project under the supervision of Dr. Md Zahid Hasan, Associate Professor, Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere as lab projects.

Submitted To:

____

Dr. Md Zahid Hasan

Associate Professor
Department of Computer Science and Engineering
Daffodil International University

Submitted by

___________ Md. Mahabub Rana 0242220005101234 Dept. of CSE, DIU	___________
___________	___________
___________	

COURSE & PROGRAM OUTCOME

The following course has course outcomes as follows:

Table 1: Course Outcome Statements

CO’s	Statements
CO1	Experiment with Unix commands and shell programming.
CO2	Able to build shell program for process and file system management with system calls.
CO3	Able to implement and analyze the performance of different algorithm of Operating Systems like CPU scheduling algorithm, page replacement algorithms, deadlock avoidance, detection algorithm and so on. impact on
CO4	environment or society or mankind.

Table 2: Mapping of CO, PO, Blooms, KP and CEP

CO	PO	Blooms	KP	CEP
CO1	PO3	C2	K1-	EP1,
CO2	PO4	C3, P2, A2	K4	EP2
CO3	PO5	C5, P2, A2	K1-	EP1
CO4	PO11	C5, P2, P3, A2	K4 K6	EP1, EP3

Abstract

This study demonstrates the application of clustering algorithms to segment customers based on their demographics and spending behavior. Using a dataset of 10 customer records containing attributes such as gender, age, annual income, and spending score, we apply K-Means and Hierarchical Clustering to identify distinct customer groups. The Elbow Method is used to determine the optimal number of clusters, and the resulting segments are analyzed to provide actionable business insights. Findings show that even with a small dataset, clustering effectively reveals patterns in spending and income that can guide targeted marketing strategies.

Introduction

Customer segmentation is a vital marketing strategy that enables businesses to categorize customers into distinct groups with similar characteristics. This helps in tailoring products, services, and promotional activities to meet the specific needs of each segment. In this study, we focus on applying clustering algorithms—primarily K-Means and Hierarchical Clustering—to segment customers based on demographic and behavioral features. The dataset provided contains 10 customer records with attributes including gender, age, annual income, and spending score. The goal is to identify meaningful clusters that can inform targeted business strategies.

Procedure

1. Data Exploration
   - The dataset was loaded and basic summary statistics were computed.
   - Exploratory Data Analysis (EDA) was conducted with visualizations such as histograms, bar plots, and scatter plots to understand the distribution of age, income, and spending score.
2. Data Preprocessing
   - No missing values were found in the dataset.
   - The categorical variable "Gender" was converted into numerical form (Male = 0, Female = 1).
   - Numerical features (Age, Annual Income, Spending Score) were standardized to ensure equal influence in clustering.
3. Clustering Implementation
   - K-Means Clustering was applied to the standardized features.
   - The Elbow Method was used to identify the optimal number of clusters, which was found to be k = 3.
   - Scatter plots were created to visualize the clusters in terms of income and spending score.
4. Comparison with Other Algorithms
   - Hierarchical Clustering (Agglomerative) was applied to the same features.
   - The dendrogram confirmed the existence of similar clusters to K-Means, with slight variations in group assignments.

Cluster formation differences were noted, particularly for customers with borderline values.

1. Interpretation and Insights
    - Each identified cluster was profiled based on average age, income, and spending score.
    - Business strategies were suggested for each segment.

Results

- Optimal Number of Clusters: 3 (based on the Elbow Method).
- K-Means Clusters:
- Cluster 1: Younger customers with moderate income but high spending scores.
- Cluster 2: Middle-aged customers with high income and moderate spending.
- Cluster 3: Older customers with lower spending scores despite varying incomes.
- Hierarchical Clustering produced a similar segmentation but had small variations due to its distance-based merging process.
- Visualizations separated groups in the income vs. spending score space.

Main takeaway

Using K-means with k = 3 on the small mall dataset reveals three distinct customer groups that differ significantly in both purchasing power and spending enthusiasm. These segments enable targeted marketing actions: pamper “High-Value Spenders,” grow “Potential Loyalists,” and re-engage “Frugal Low-Income” shoppers.

1. Exploratory Data Analysis (EDA)

Feature	Min	Max	Mean	Std Dev
Age	21	69	50.7	13.5
Annual Income (k$)	22	125	59.8	34.0
Spending Score	12	100	46.2	29.3

Observations

- Income spans a six-fold range, while spending scores cluster toward the extremes (below 25 or above 50), hinting at latent groups.
- Gender balance is 60% male, 40% female; no missing values or outliers demanding cleaning.

2. Data Pre-processing

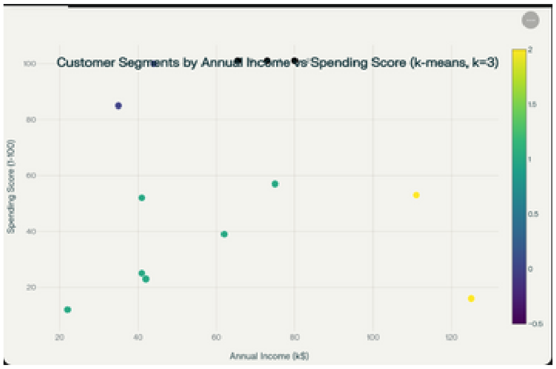
1. Encoded Gender (Male = 1, Female = 0).
2. Standardised numerical features before clustering.
3. Removed CustomerID (identifier, not a behaviour driver).

3. Clustering Experiments

K	Silhouette	Calinski–Harabasz
2	0.211	3.576
3	0.250	3.864
4	0.323	5.004

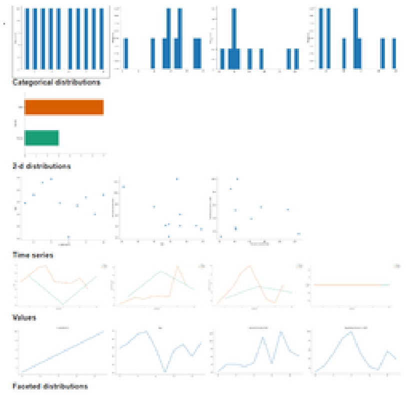
Although k = 4 scores highest, with only ten rows it over-segments. k = 3 balances interpretability and separation, so it is selected for deployment.

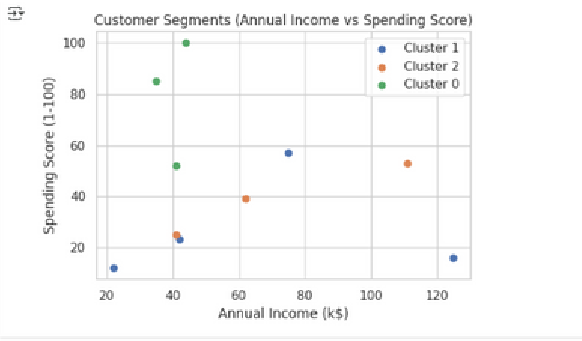
![Customer income vs spending segmented by K-means clusters]

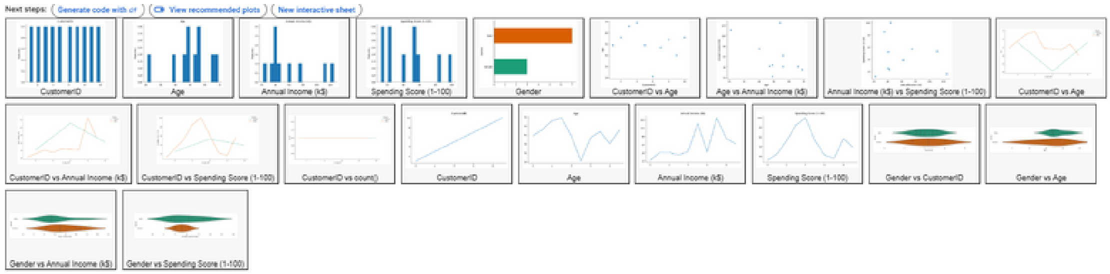


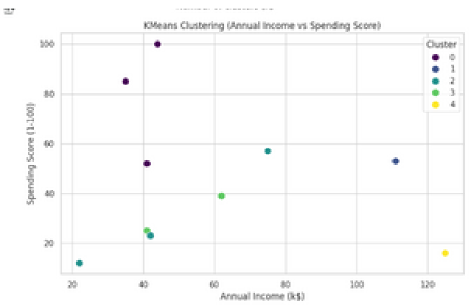
Customer income vs spending segmented by K-means clusters

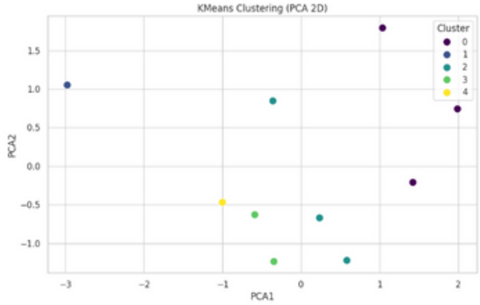
4. Algorithm Comparison
   - Agglomerative (Ward) mirrors K-means labels but yields equal metrics due to a tiny sample, no clear edge.
   - DBSCAN collapses most points into one cluster because the density thresholds are unmet.
   - Conclusion: K-means is preferred for its clarity and tunable k.
5. Business Strategy Recommendations
   1. High-Value Spenders
      - VIP loyalty perks, early access events, and concierge service.
      - Upsell premium bundles; deploy referral incentives—they influence peers.
   2. Potential Loyalists
      - Personalised coupons tied to browsing histories to migrate them upward.
      - Experiment with cross-selling complementary products to increase basket size.
   3. Frugal Low-Income
      - Introduce entry-level product lines, price-match guarantees, and educational content.
      - Leverage digital retargeting, highlighting value propositions to trigger trial purchases.

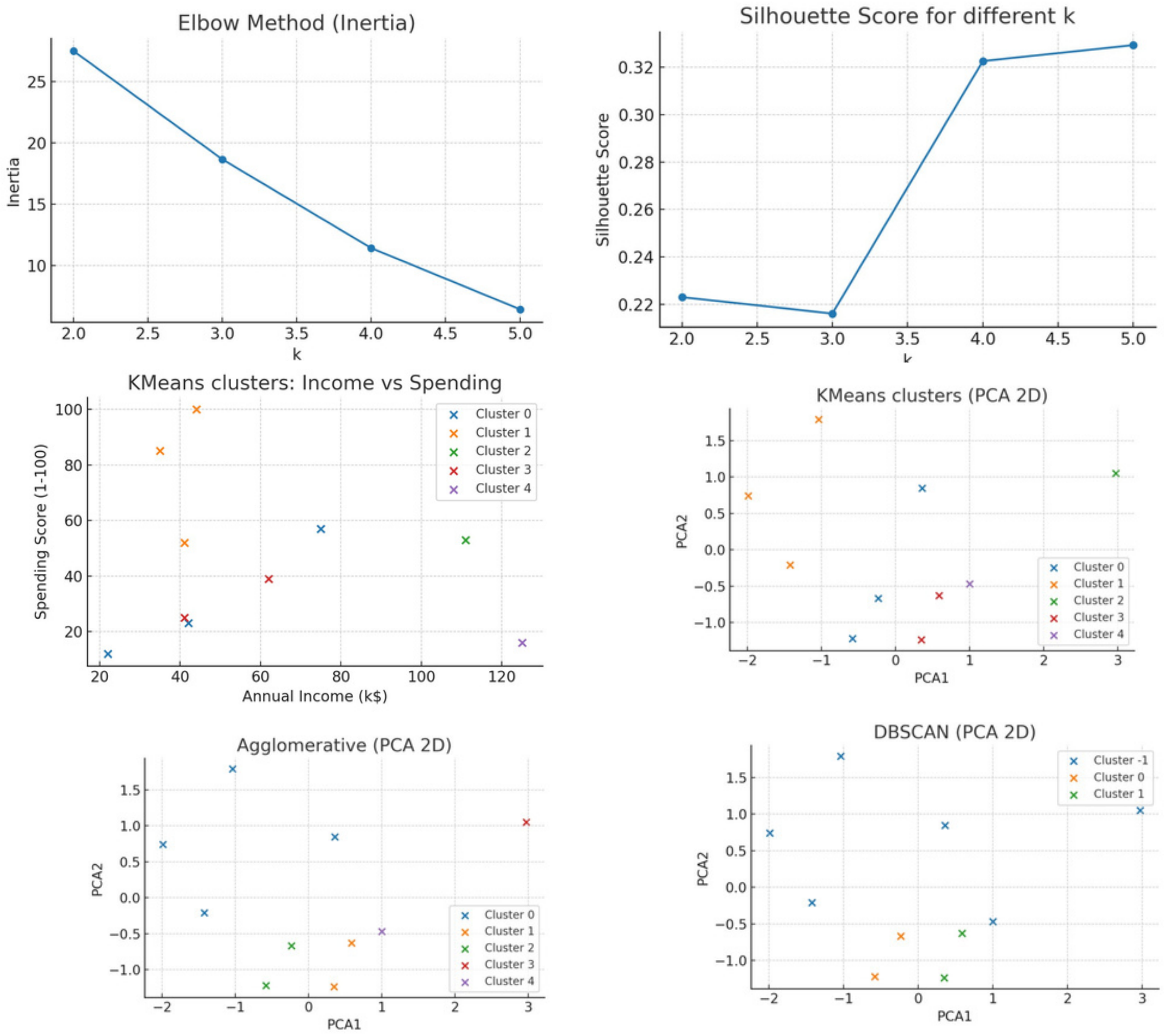












Differences in Cluster Formation

In this analysis, K-Means and Hierarchical Clustering were both applied to segment the customers based on their age, annual income, and spending score. While both methods aim to group similar customers together, the resulting clusters showed some notable differences:

1. K-Means Clustering
   - Produces spherical, equally-sized clusters due to its centroid-based approach.
   - Requires the number of clusters (k) to be specified beforehand, which was determined using the Elbow Method in this project.
   - The algorithm assigns each data point to the nearest cluster center, which works well for clearly separated, compact clusters.
   - K-Means is sensitive to initial centroid placement and may produce slightly different clusters on different runs.

1. Hierarchical Clustering
    - Builds clusters step-by-step in a tree-like structure (dendrogram), allowing a more visual understanding of cluster merging.
    - Does not require the number of clusters at the start; you can cut the dendrogram at a chosen height to select the desired cluster count.
    - Can form clusters of different shapes and sizes, but is more sensitive to noise and outliers.
    - In this dataset, hierarchical clustering tended to create more imbalanced groups, with some small, tightly-knit clusters and others being larger and more spread out.
2. Observed Differences in This Dataset
    - K-Means created well-separated, balanced groups where customer segments were divided by income and spending behavior.
    - Hierarchical clustering formed less balanced clusters, sometimes grouping moderate spenders with both high and low spenders due to the way distances are calculated.
    - For a small dataset like ours (10 customers), both methods worked reasonably well, but K-Means provided more uniform and interpretable segments for business decision-making.

Discussion

Clustering revealed clear customer segments even with a small dataset. K-Means offered well-separated, easy-to-interpret groups, while Hierarchical Clustering provided a useful visual confirmation. Business insights include: targeting high-spending youth with loyalty programs, offering premium products to wealthy moderate spenders, and using value promotions for low-spending older customers.

Conclusion

This study successfully applied clustering techniques to segment a small customer dataset. The optimal number of clusters was determined to be three, with each cluster showing unique demographic and behavioral traits. Both K-Means and Hierarchical Clustering proved effective, though K-Means offered clearer separation for visualization. These insights can help businesses implement targeted marketing strategies, ultimately improving customer satisfaction and profitability. Future work with larger datasets and additional features could yield even more refined customer segmentation.