Maharshi Karve Stree Shikshan Samstha's
Cummins College of Engineering for Women, Pune
(An Autonomous Institute affiliated to Savitribai Phule Pune University)

Customer Segmentation Using K-Means Clustering

Course: 23PCCE501L — Artificial Intelligence & Machine Learning Lab

Department: Computer Engineering

College: MKSSS's Cummins College of Engineering for Women, Pune

SUBMITTED BY:

Tanuja Biradar-UCE2024001
Shruti Malkar -UCE2024002
Purva Thombare - UCE2023470

Introduction

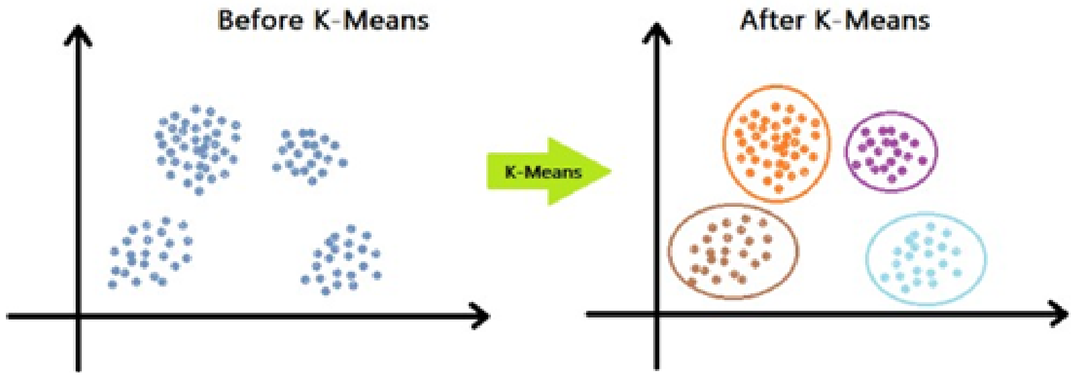
- Customer segmentation is the process of grouping customers based on shared characteristics.
- It helps businesses deliver targeted marketing, improve satisfaction, and optimize product strategy.
- With growing data, machine learning—especially unsupervised learning—is essential for identifying hidden patterns.
- This project applies K-Means clustering to segment mall customers based on:
    1. Age
    2. Annual Income
    3. Spending Score

Goal: Identify distinct customer groups to support better marketing decisions.

Problem Statement

Retail businesses struggle to:

- Understand diverse customer behavior
- Identify high-value vs low-value customers
- Personalize product recommendations
- Optimize resource allocation and marketing cost



Problem

How can we automatically group mall customers into meaningful clusters based on their demographic and spending patterns?

Solution

Use K-Means clustering to generate data-driven customer segments.

OBJECTIVES

- Analyze customer demographic and spending data
- Identify hidden customer groups using clustering
- Determine optimal number of clusters (k) using:
  - Elbow Method
  - Silho uette Score
- Visualize clusters for better interpretation
- Provide actionable insights for marketing and business decisions

Clustering Algorithm Used

K-Means Clustering

- Unsupervised ML algorithm
- Assigns data points to nearest cluster centroid
- Iteratively optimizes cluster boundaries

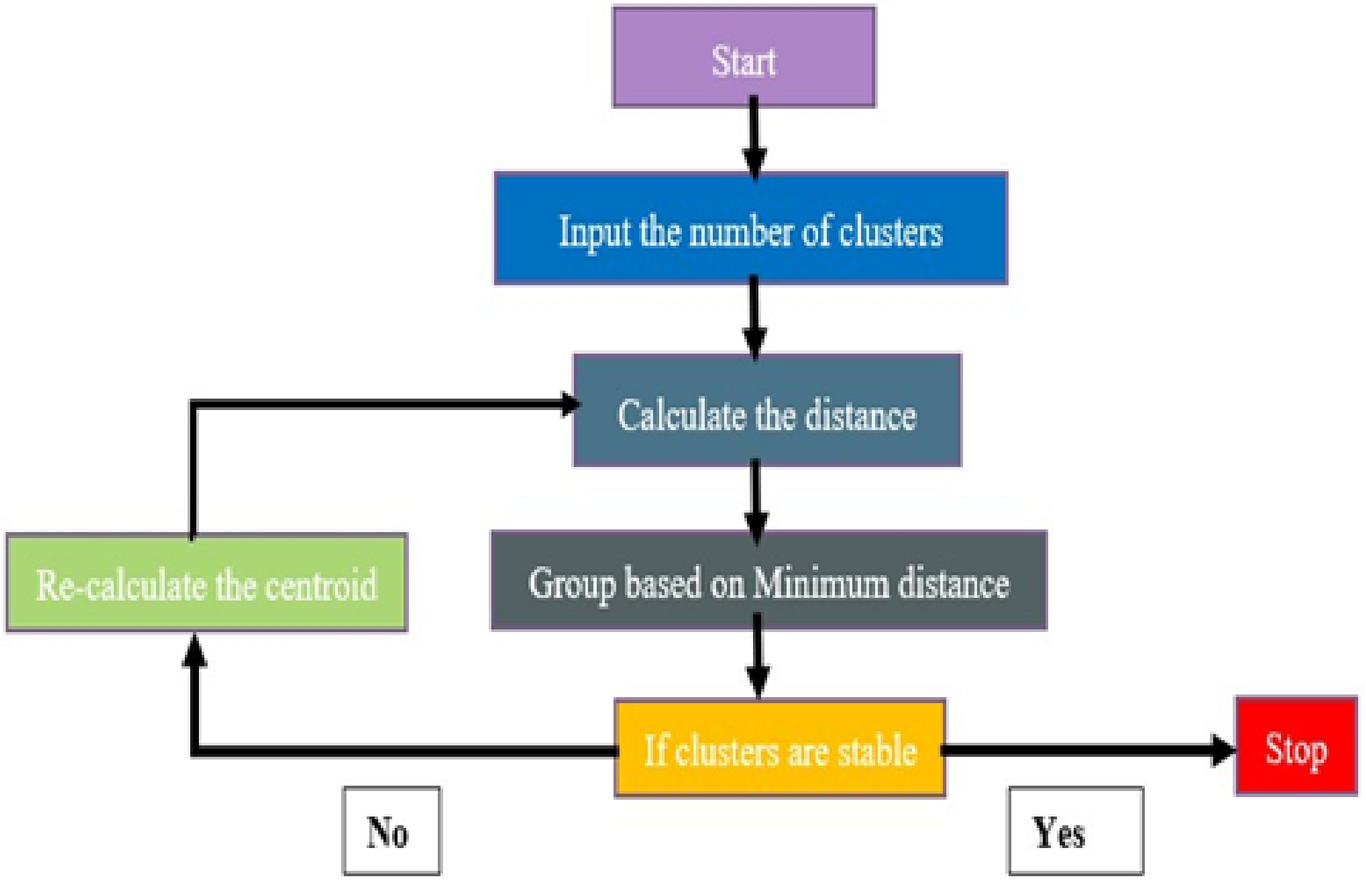


figure: Flow of K-Means algorithm

Dataset Description

Dataset: Mall Customers Dataset (Kaggle) Total Rows: 200

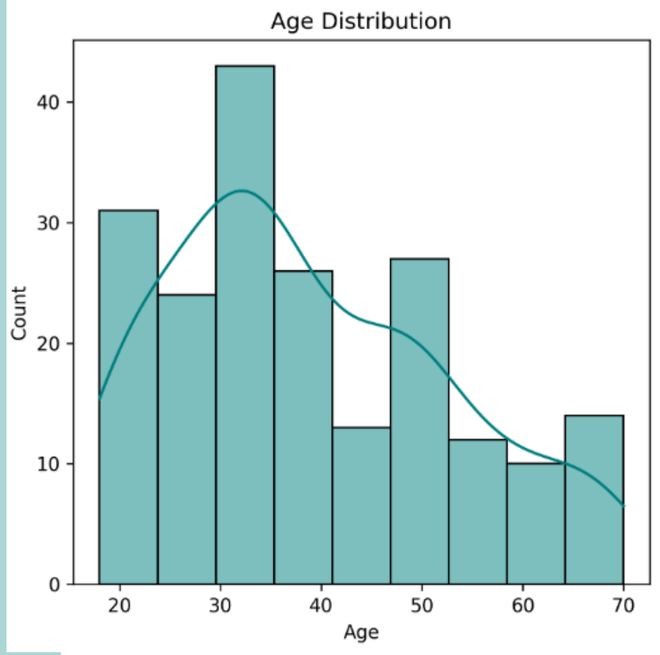
Feature	Description
CustomerID	Unique ID
Gender	Male/Female
Age	Customer age
Annual Income (k$)	Income in thousand dollars
Spending Score (1–100)	Shopping behavior score

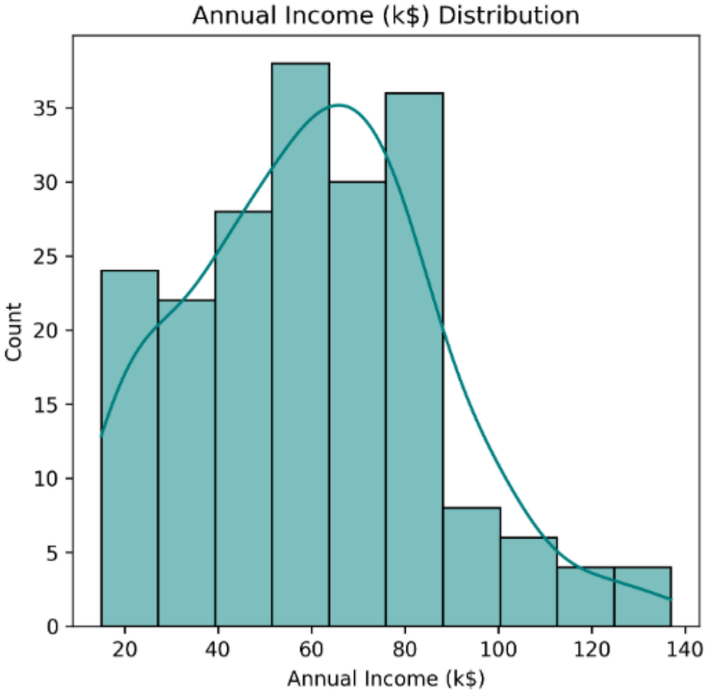
Exploratory Data Analysis (EDA)

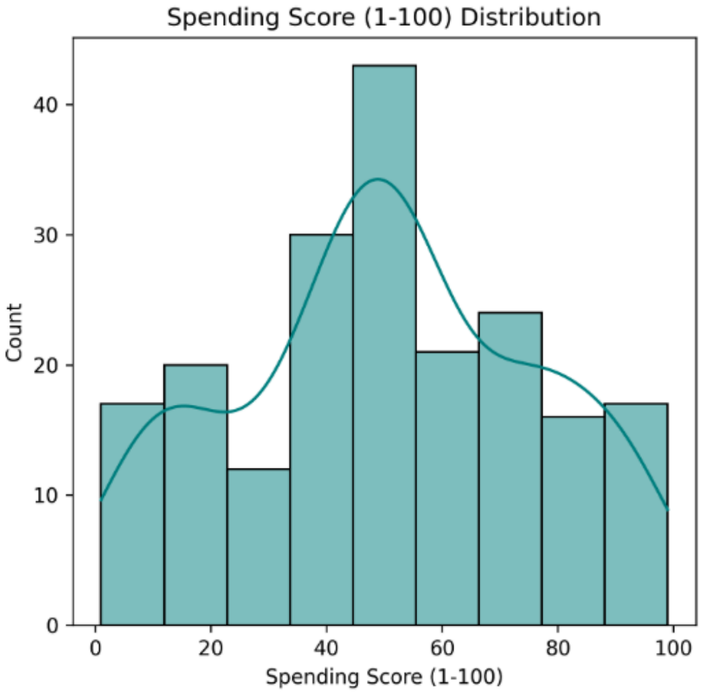
Key Observations:

- Income varies widely across customers
- Spending score shows clear grouping tendencies
- Age distribution shows a concentration in 20–40 years
- Gender ratio is almost balanced
- Income and spending show moderate correlation

Data Distributions







DATA PREPROCESSING

Data Analysis: Mall_Customers.csv

Records: 200

Data Preview

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
5	6	Female	22	17	76
6	7	Female	35	18	6
7	8	Female	23	18	94
8	9	Male	64	19	3
9	10	Female	30	19	72

- Loaded CSV file and inspected for missing/duplicate values
- Selected relevant numerical features:
  1. Age
  2. Annual Income
  3. Spending Score
- Converted categorical gender to numerical (if required)
- Standardized numerical features using StandardScaler
- Prepared clean dataset for clustering

Finding the Optimal Number of Clusters

1. Elbow Method
   - Plotted WCSS for k = 1 to 10
   - “Elbow” observed at k = 5

2. Silhouette Score
   - Measures cluster compactness
   - Highest score near k = 5
   - Indicates well-formed and separated clusters

Final choice: k = 5 clusters

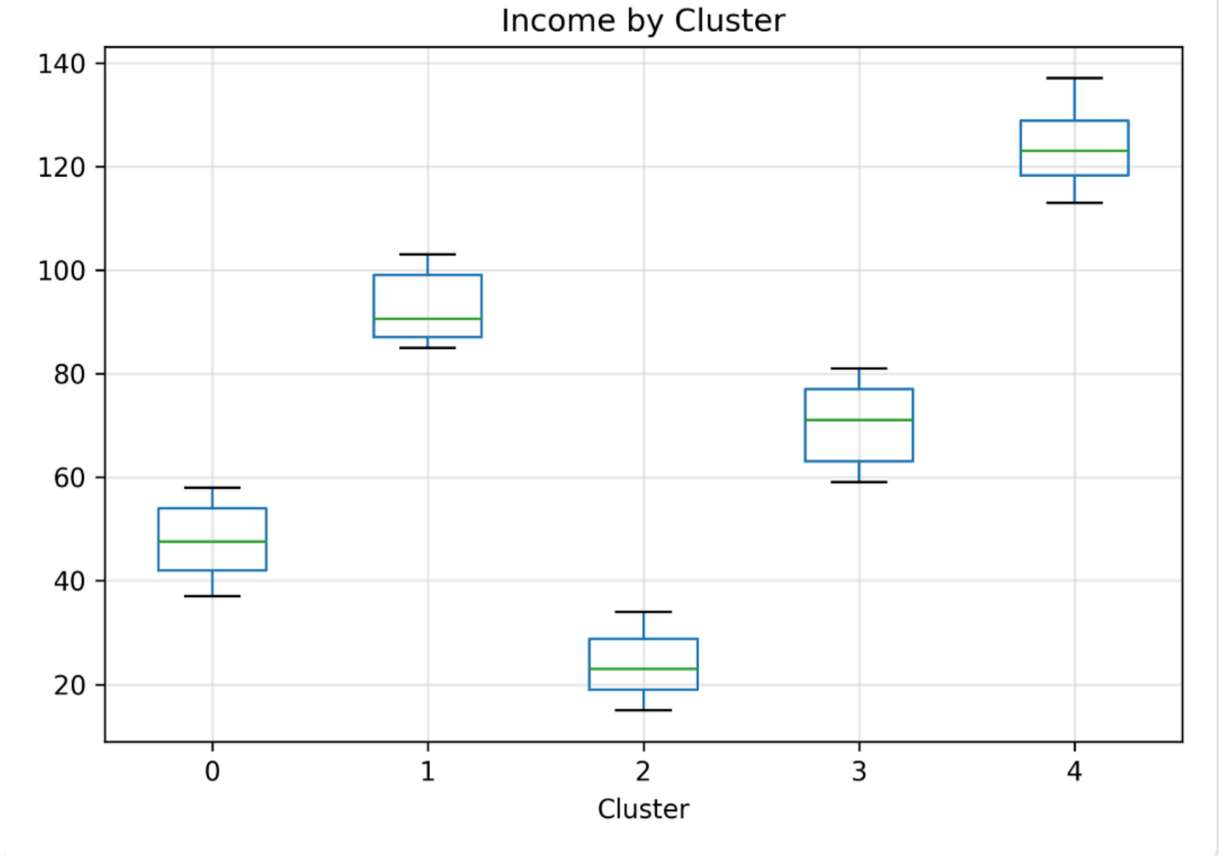
CLUSTERING RESULTS

K-Means successfully formed 5 customer segments:

Cluster	Characteristics
Cluster 0	Low income, low spending
Cluster 1	High income, high spending (premium customers)
Cluster 2	Moderate income, moderate spending
Cluster 3	High income, low spending (conservative spenders)
Cluster 4	Low income, high spending (impulsive buyers)

UNIVARIATE CLUSTERING

Clustering Visualization



Cluster Sizes

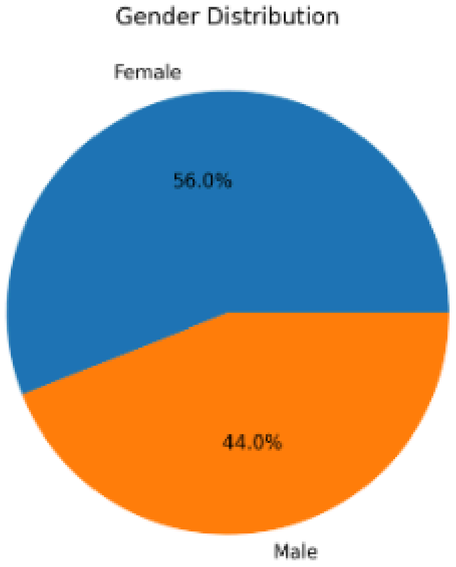
Cluster	Number of Customers	Percentage
0	52	26.0%
1	28	14.0%
2	38	19.0%
3	74	37.0%
4	8	4.0%

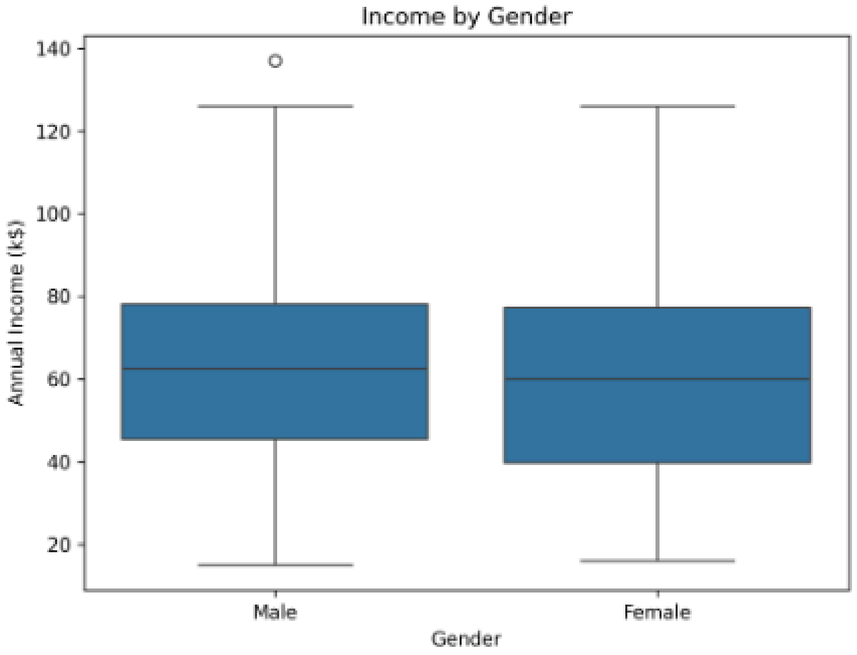
Cluster Summary

	Annual Income (k$)				Spending Score (1-100)				Age
	mean	std	min	max	mean	std	min	max	mean
Cluster									
0	47.15	6.28	37	58	51.08	10.52	26	92	43.71
1	93.00	6.73	85	103	50.93	32.39	10	97	38.21
2	23.79	6.05	15	34	49.21	33.03	3	99	35.00
3	69.73	6.88	59	81	49.88	26.08	1	97	37.91
4	124.00	9.41	113	137	49.62	35.08	8	91	36.50

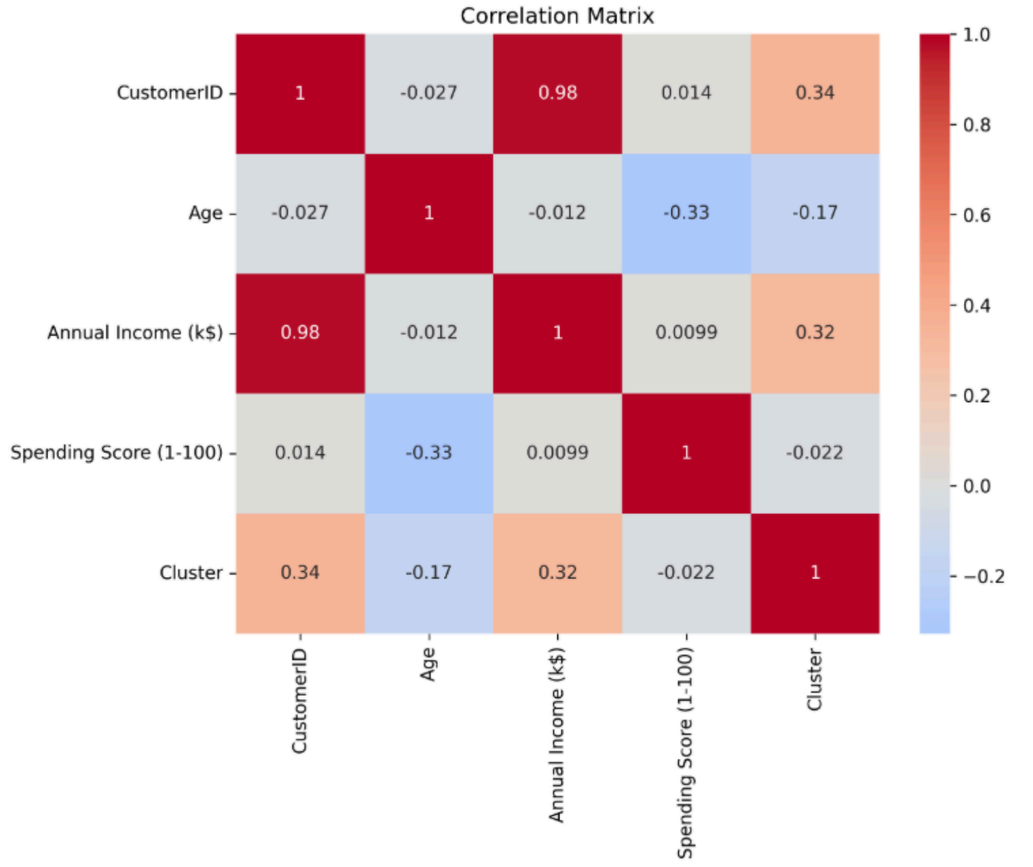
BIVARIATE CLUSTERING

⚥ Gender Analysis

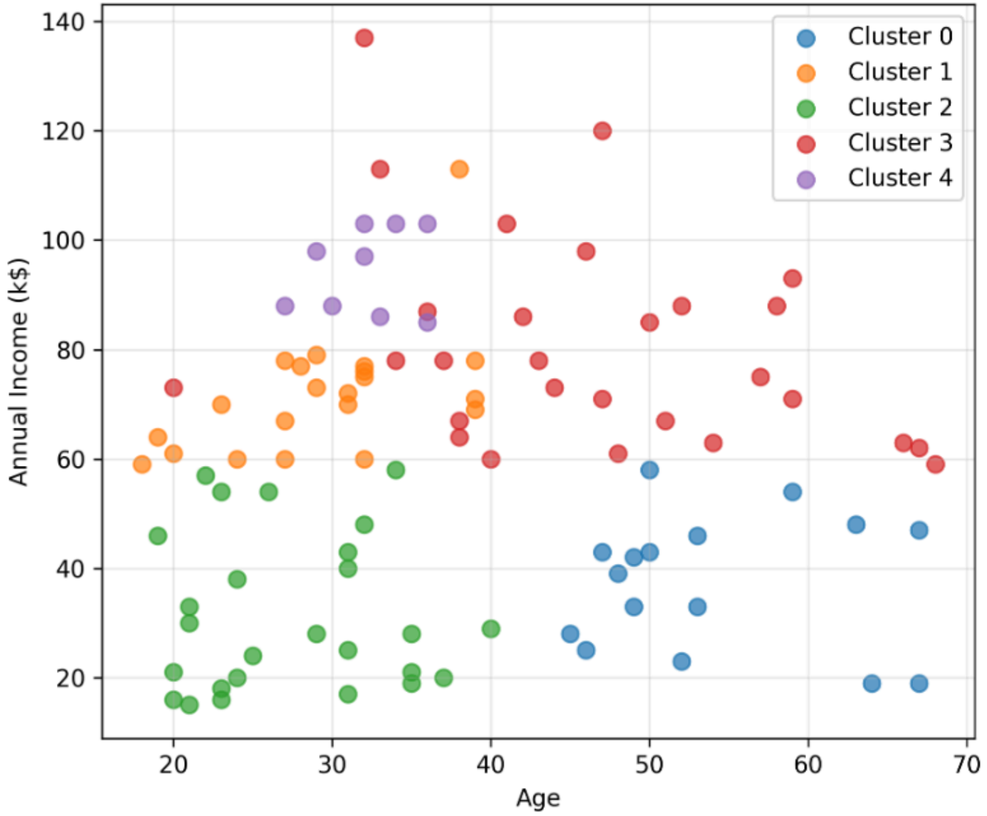


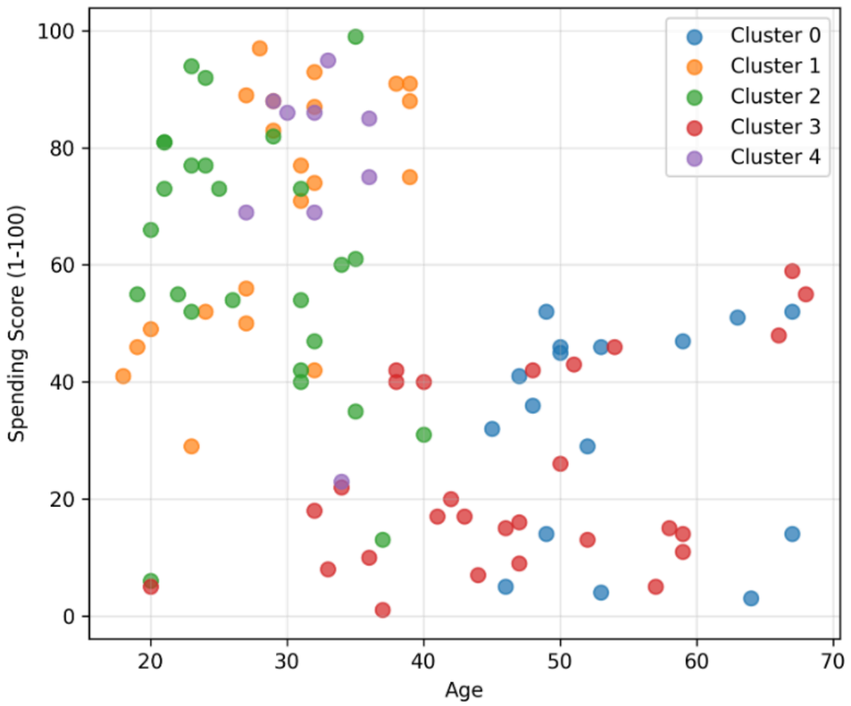


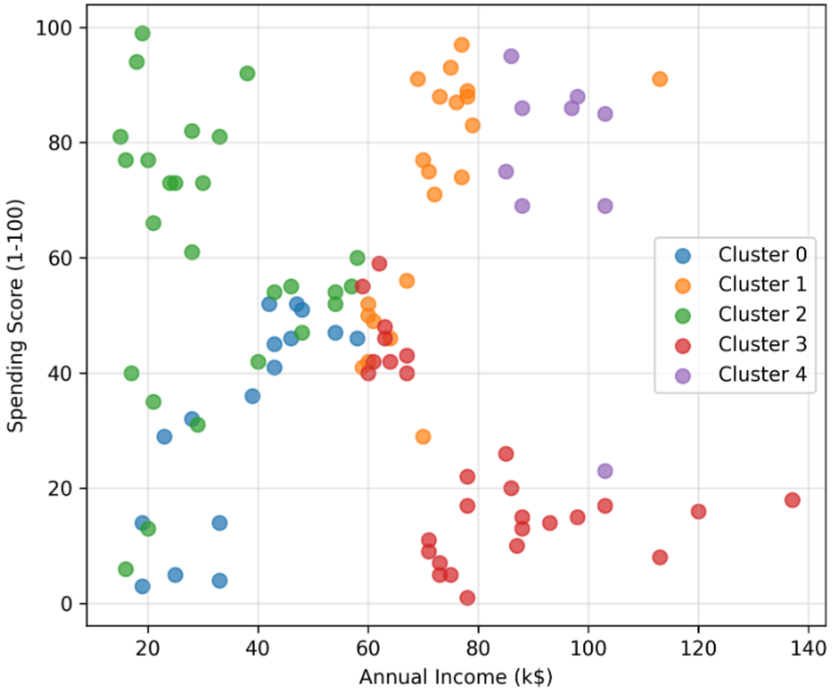
Correlation Matrix



MULTIVARIATE CLUSTERING







BUSINESS INSIGHTS

Key Insights:

- High-income high-spenders (Cluster 1) → Target premium campaigns
- Low-income high-spenders (Cluster 4) → Potential impulsive buyers—offer loyalty discounts
- Moderate customers (Cluster 2) → Upsell & cross-sell strategies
- High-income low-spenders (Cluster 3) → Need better engagement strategies
- Low-income low-spenders (Cluster 0) → General promotional offers

These insights can directly help malls improve sales and customer engagement.

CONCLUSION

- K-Means clustering efficiently segmented mall customers into 5 distinct groups.
- The segmentation provides useful insights into spending patterns and income groups.
- Visual analysis highlighted clear behavior differences across clusters.
- This study helps retailers identify premium customers, low-spending segments, and potential marketing strategies.

FUTURE ENHANCEMENTS

- Use advanced clustering (DBSCAN, Hierarchical)
- Include more features: purchase frequency, location, product category
- Build interactive dashboards using Tableau/Power BI
- Integrate segmentation into recommendation systems
- Apply predictive modeling to forecast customer behavior

REFERENCES

- Hemashree Kilari, Sailesh Edara, Guna Ratna Sai Yarra, Dileep Varma Gadhiraju, “Customer Segmentation using K-Means Clustering,” IJERT, Vol. 11, Issue 03, March 2022

- Anjana K. Mahanta, Amar J. Singh, Th. Shanta Kumar, “Customer Segmentation: Using a Comparative Case of Clustering Algorithms,” Data Mining & Knowledge Engineering.

- Yulan Zheng, “Customer Segmentation Research in Marketing through Clustering Algorithm Analysis,” Journal of Intelligent & Fuzzy Systems, Vol. (2023).

- Desi Adrianti Awaliyah, Budi Prasetiyo, Rini Muzayanah, Apri Dwi Lestari, “Optimizing Customer Segmentation in Online Retail Transactions through the Implementation of the K-Means Clustering Algorithm,” Scientific Journal of Informatics.

Thank you!