# Report of Clustering Methods

Dataset - Mall Customer

Subject - Machine Learning

## Objective

The objective of this lab is to gain hands-on experience with unsupervised learning by applying three clustering algorithms: **K-Means**, **Agglomerative Hierarchical Clustering**, and **DBSCAN**. We aim to identify customer segments using the *Mall Customers* dataset and interpret them from a business perspective.

—

## Dataset

The dataset (`Mall_Customers.csv`) contains 200 entries with the following columns:

- **CustomerID**: Unique identifier for each customer.
- **Gender**: Male or Female.
- **Age**: Age of the customer.
- **Annual Income (k$)**: Customer's annual income in thousands of dollars.
- **Spending Score (1–100)**: A score assigned based on customer behavior.

For clustering, only **Annual Income** and **Spending Score** were used, as per lab instructions.

—

## Data Exploration and Visualization

After loading the dataset into a Pandas DataFrame:

- The data contained no missing values.
- A scatter plot of Annual Income vs Spending Score showed distinct dense regions, hinting at possible clusters.

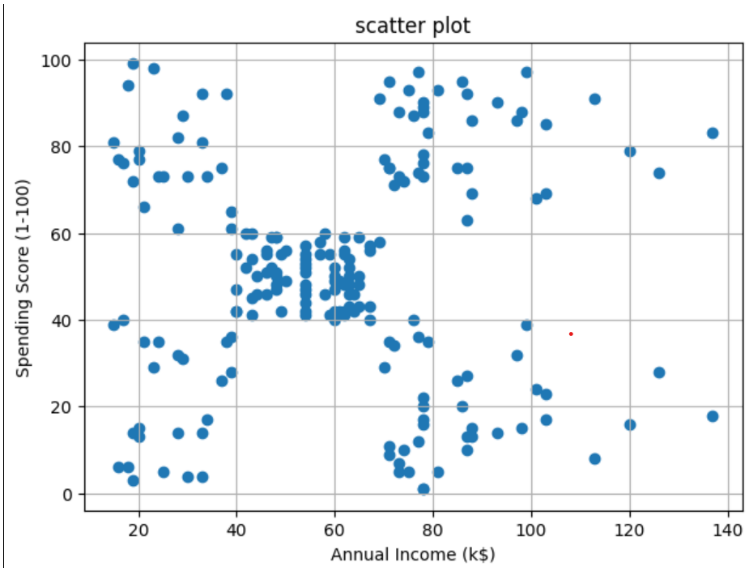

Figure 1: Scatter plot of Annual Income vs Spending Score

—

# K-Means Clustering

## Elbow Method

The Elbow Method was applied by computing WCSS for $k = 1$ to $k = 10$. A clear "elbow" appeared at $k = 5$, suggesting five clusters.

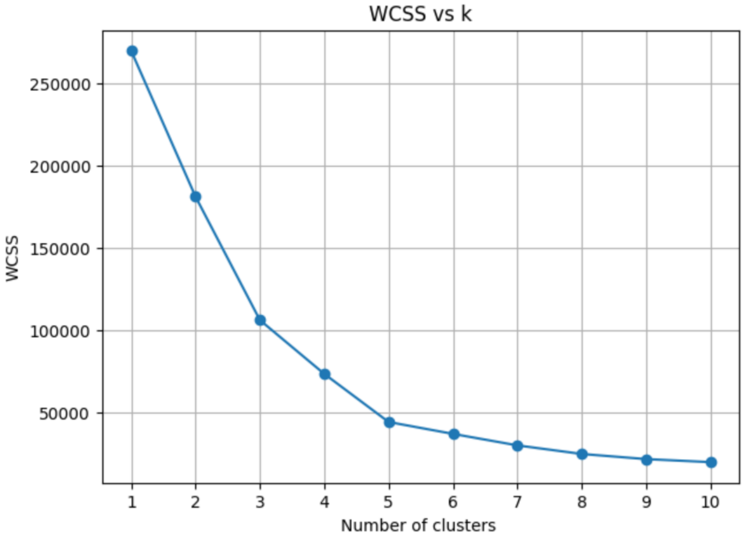

Figure 2: Elbow Method for K-Means

## Final Clustering

K-Means was applied with $k = 5$. The resulting clusters clearly separated customers into distinct spending/income groups.

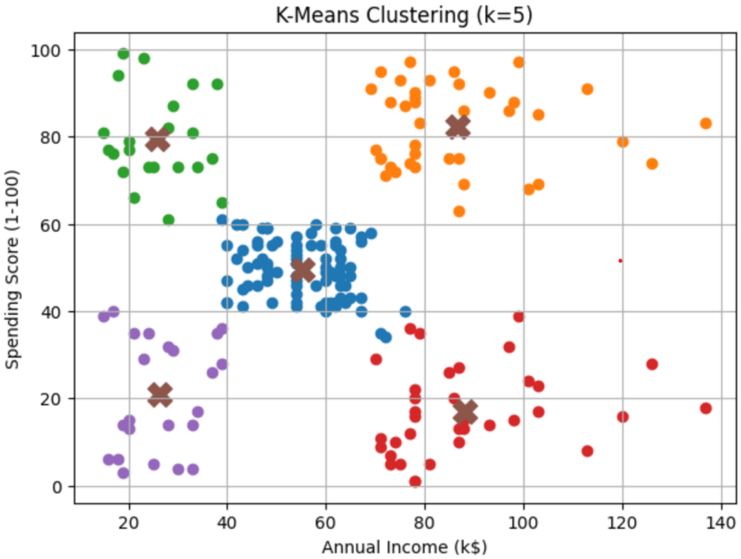

Figure 3: K-Means clustering with $k = 5$ (centroids marked with X)

**Interpretation:** The five clusters represented:

1. Low income, low spending.
2. High income, high spending (premium customers).
3. Medium income, medium spending.
4. Low income, high spending (value-driven shoppers).
5. High income, low spending (reluctant spenders).

—

# Agglomerative Hierarchical Clustering

## Dendrogram

A dendrogram using Ward linkage suggested around 5 clusters, consistent with the K-Means result.

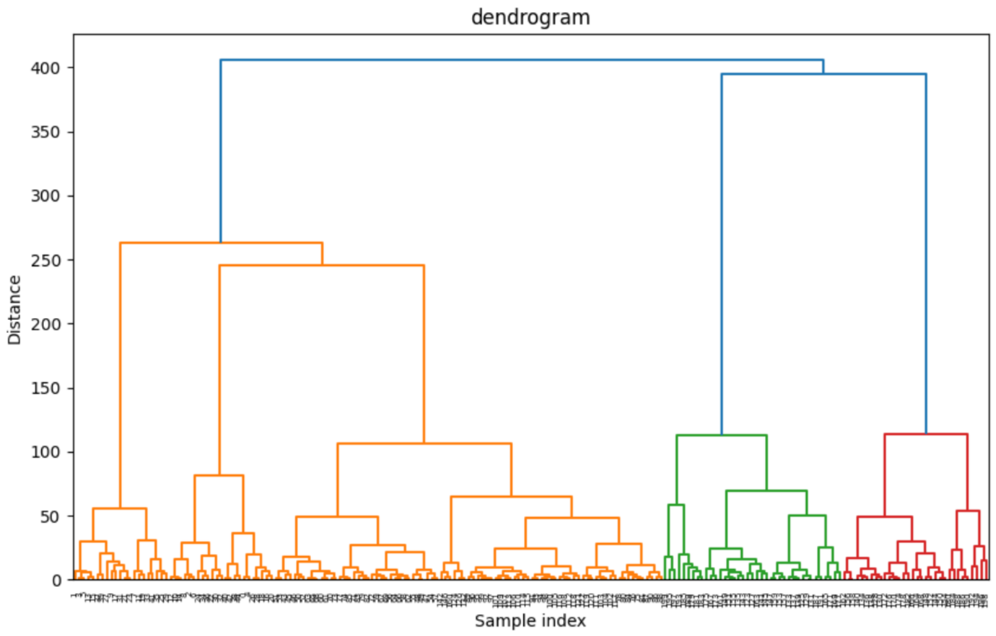

Figure 4: Dendrogram using Ward linkage

## Final Clustering

Agglomerative clustering with 5 clusters gave results similar to K-Means, with only minor differences at the cluster boundaries.

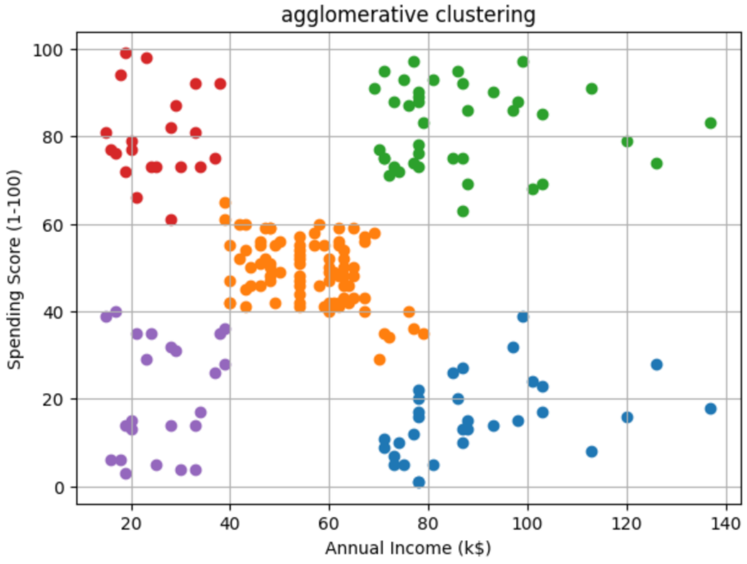

Figure 5: Agglomerative Clustering (5 clusters)

# DBSCAN

DBSCAN was applied with different parameters:

- **eps = 5, min_samples = 5**: Produced 4 clusters and some noise points.
- **eps = 3, min_samples = 5**: Marked many points as noise due to stricter neighborhood size.
- **eps = 6, min_samples = 4**: Fewer noise points, larger clusters.

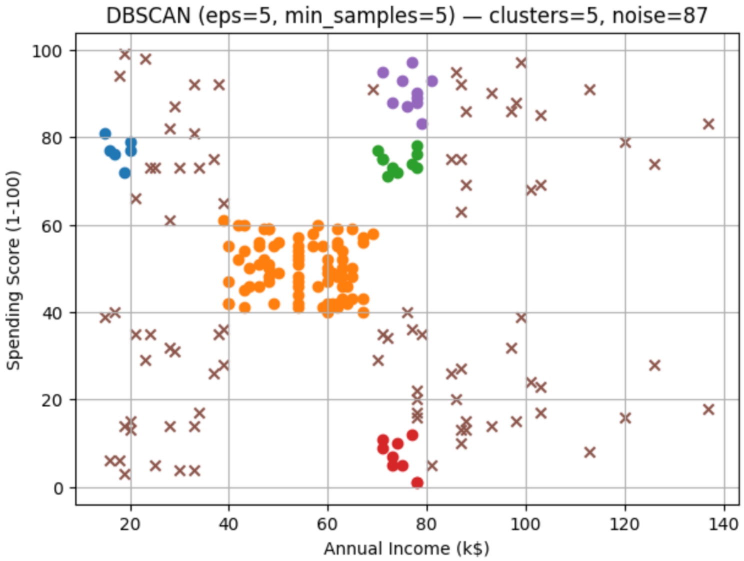

Figure 6: DBSCAN clustering with eps=5, min_samples=5

**Observation:** DBSCAN did not perform as consistently as K-Means or Agglomerative for this dataset, because the customer groups are fairly compact and spherical.

—

# Analysis and Questions

## 1. Optimal Clusters

- K-Means: 5 clusters (Elbow Method). - Hierarchical: 5 clusters (Dendrogram). - DBSCAN: Cluster count varied depending on parameters.

## 2. Cluster Comparison

K-Means and Hierarchical produced similar, interpretable clusters. DBSCAN produced fewer clusters and marked noise points.

### 3. DBSCAN Performance

DBSCAN identified outliers, but was sensitive to parameter choice. Its results were less meaningful compared to K-Means and Hierarchical.

### 4. Algorithm Suitability

For this dataset, K-Means and Agglomerative (Ward) are most suitable because the clusters are compact. DBSCAN is less suitable here.

### 5. Real-World Application

A mall could use these clusters for marketing:

- **High income, low spending:** Target with exclusive offers and personalized promotions.
- **Low income, high spending:** Offer budget-friendly deals and loyalty rewards.
- **High income, high spending:** Maintain premium relationships with VIP perks.

—

## Conclusion

- K-Means and Agglomerative clustering both identified five meaningful customer segments.
- DBSCAN struggled due to the dataset's structure but is valuable for irregular clusters or outlier detection.
- Customer segmentation can directly support targeted marketing strategies and personalized campaigns and premium experience for the cluster who earn high but spend low.