Clustering Methods Applied to Mall Customers Dataset

Prisha Khalasi[1]
[1] DAU, Gandhinagar.

Keywords: **Keywords:** Clustering, K-Means, Hierarchical Clustering, DBSCAN, Customer Segmentation.

Abstract: This paper applies and compares three clustering methods—K-Means, Agglomerative Hierarchical Clustering, and DBSCAN—on the Mall Customers dataset. We evaluate clustering performance, determine the optimal number of clusters, and discuss real-world applications in customer segmentation to improve targeted marketing strategies.

1 Introduction

Unsupervised learning methods are vital in discovering hidden data structures without labeled outcomes. Clustering algorithms group data based on similarity, aiding in customer segmentation to inform business decisions. This work implements K-Means, Agglomerative Hierarchical Clustering, and DBSCAN on a mall customers' dataset focusing on annual income and spending score features.

2 Data and Methods

The Mall Customers dataset includes 200 customers with attributes including annual income (in k$) and spending score (1–100). Only these two features are used.

2.1 Clustering Algorithms

K-Means Clustering partitions data into clusters minimizing within-cluster variance:

$$J = \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \mu_i\|^2$$

Agglomerative Hierarchical Clustering merges clusters based on Ward's linkage:

$$d(A, B) = \frac{|A||B|}{|A| + |B|} \|\mu_A - \mu_B\|^2$$

DBSCAN identifies dense regions with parameters ϵ and minPts, labeling points as core, border, or noise.

2.2 Optimal k Selection

K-Means optimal clusters determined by the Elbow Method, plotting inertia over k:

$$Inertia(k) = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mu_i\|^2$$

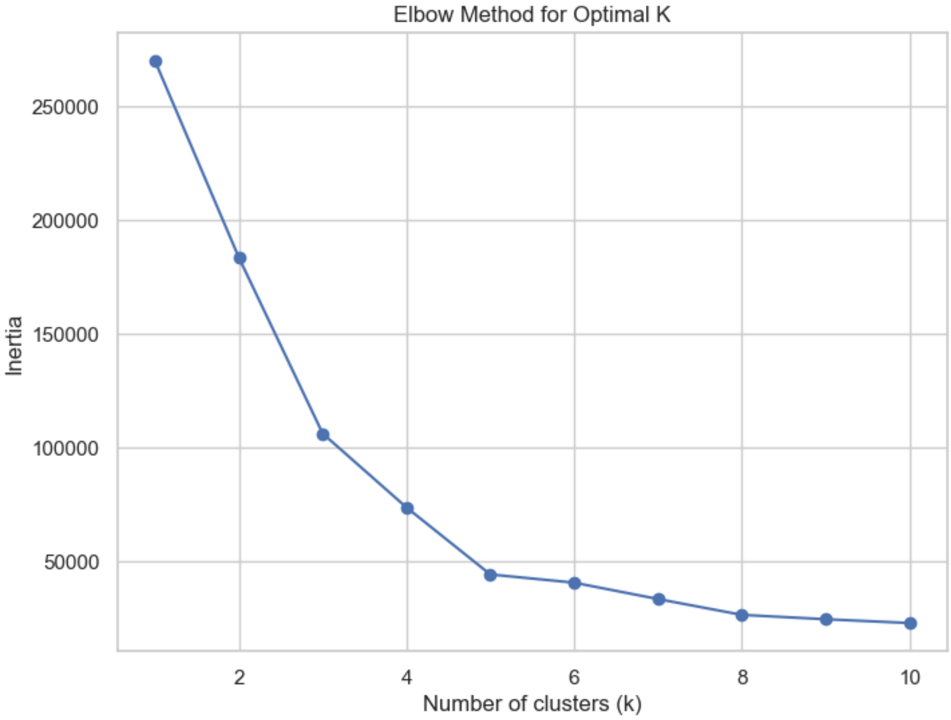

Figure 1: Elbow Method for Optimal K

3 Results

3.1 Elbow Plot

3.2 Data Visualization

3.3 Hierarchical Clustering

3.4 Clustering Results

4 Discussion

4.1 Real-World Application of Customer Segments

The clustering reveals distinct segments:

High Income, Low Spending Customers represent untapped potential. Strategies include exclusive offers and loyalty programs to increase spending.

Low Income, High Spending Customers may benefit from financing options to support purchasing power.

Average Income and Spending Customers can be targeted with seasonal promotions and upselling. This segmentation supports efficient marketing allocation and enhances revenue.

5 Conclusion

Clustering the Mall Customers dataset with K-Means, Hierarchical, and DBSCAN reveals meaningful customer segments, facilitating targeted marketing. K-Means and Agglomerative produce consistent results, while DBSCAN effectively identifies noise/outliers.

Article version: h CET

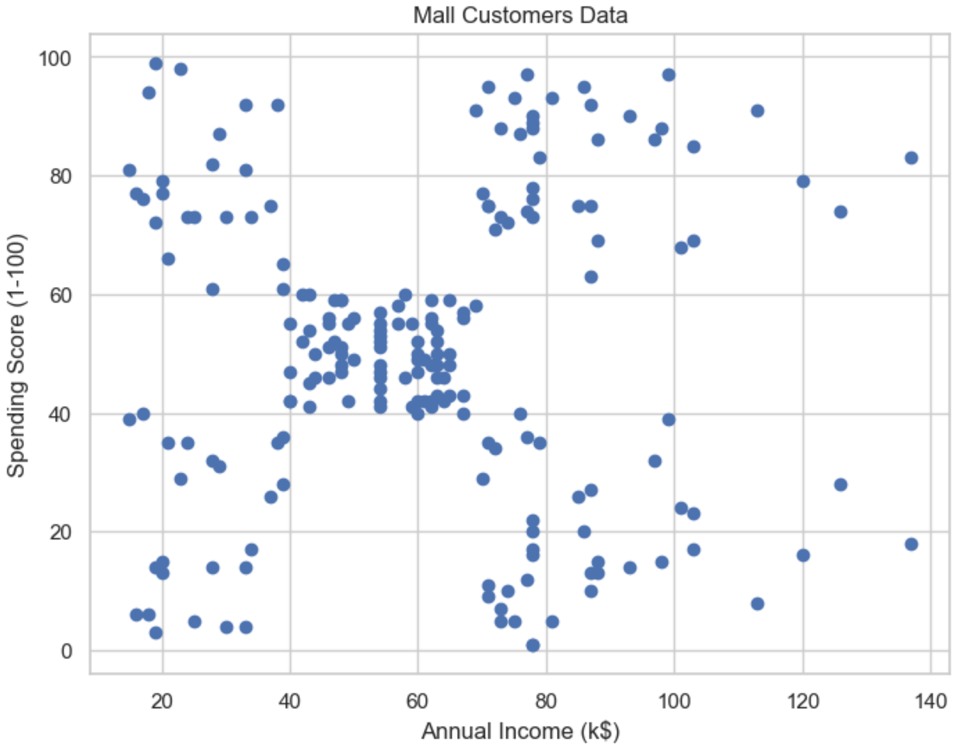

Figure 2: Mall Customers Data: Annual Income vs. Spending Score

Conflict of interests

None declared.

Acknowledgments

The author thanks the CIBB community for fruitful discussions.

Funding

No external funding was received.

Availability of data and software code

Data and code are available on request.

References

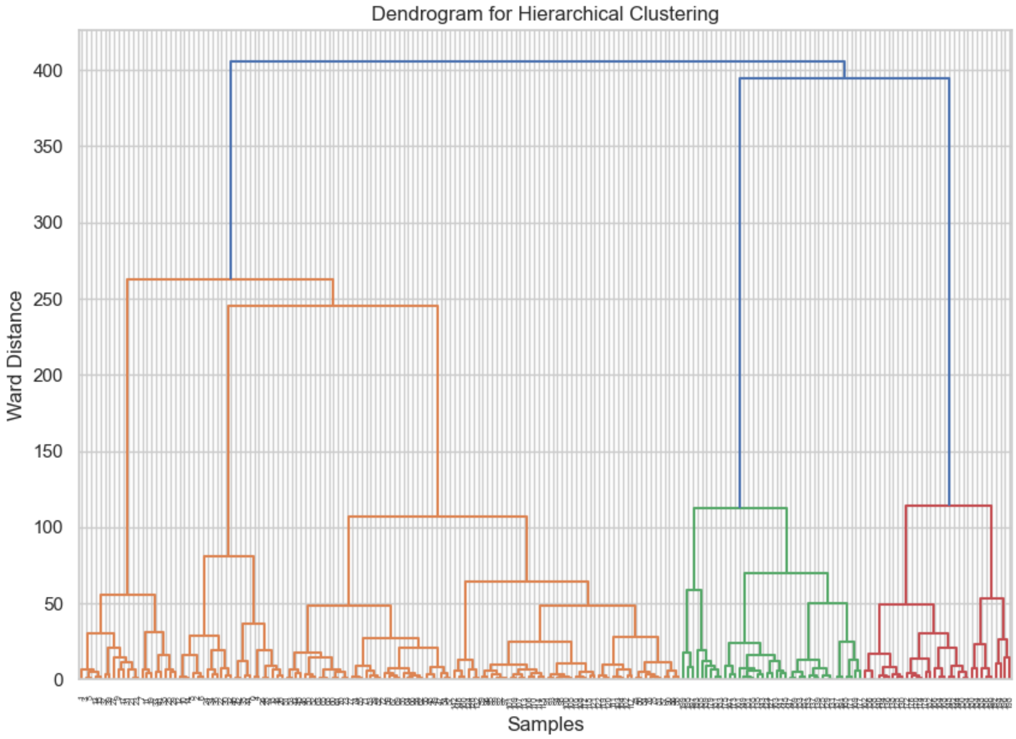

Figure 3: Dendrogram for Hierarchical Clustering

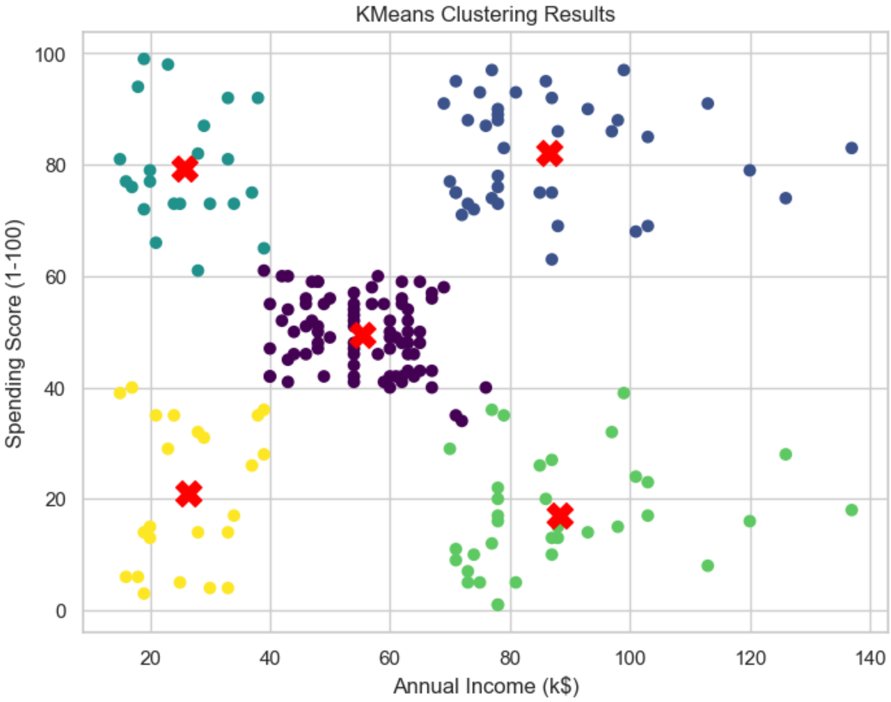

Figure 4: K-Means Clustering Results

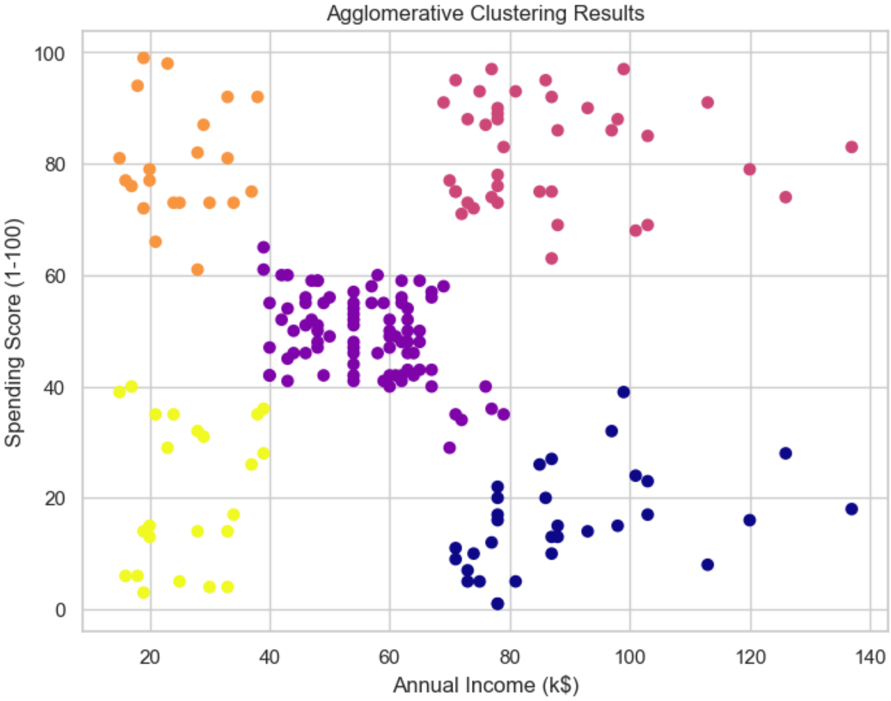

Figure 5: Agglomerative Clustering Results

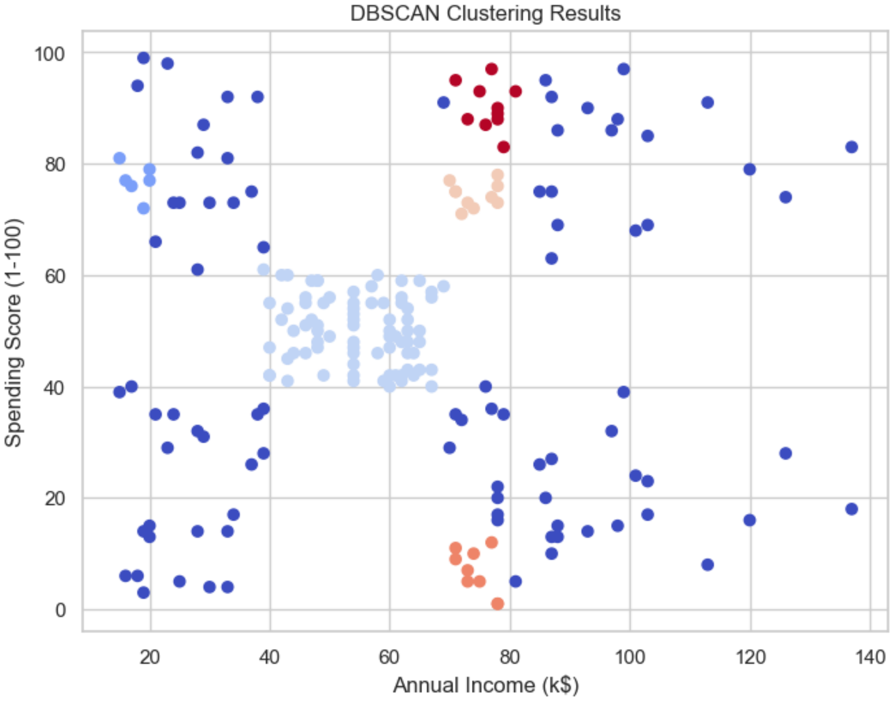

Figure 6: DBSCAN Clustering Results