 AFRICAN MASTERS OF MACHINE INTELLIGENCE
(AMMI)

K-Means Clustering Analysis on Mall Customers Dataset

Njombe Ejah Julius Dilan

May 28, 2025

Abstract

This report presents an analysis of customer segmentation using K-Means clustering applied to the *Mall Customers* dataset from Kaggle. Various initialization strategies were implemented and evaluated using the Elbow Method and Silhouette Score. The goal is to identify optimal clusters of customers based on features from the dataset such as Annual Income and Spending Score.

1 Introduction

Customer segmentation is an important task in marketing that allows businesses to provide their services based on customer characteristics. K-Means clustering is one of the most widely used unsupervised learning techniques for this purpose [1]. This report explores the use of K-Means with multiple initialization methods that include:

- Random Initialization

- K-Means++ Initialization

- Max-Distance Initialization

2 Dataset Description and Data Exploration

The **Mall Customers** dataset contains **200** observations and **5** features, with no missing values.
The features are:

- **CustomerID**

- **Gender**

- **Age**

- **Annual Income (k$)**

- **Spending Score (1–100)**

For clustering purposes, we used the scaled values of **Annual Income** and **Spending Score**.

The figure below shows the distributions of the three main numerical features in the dataset:
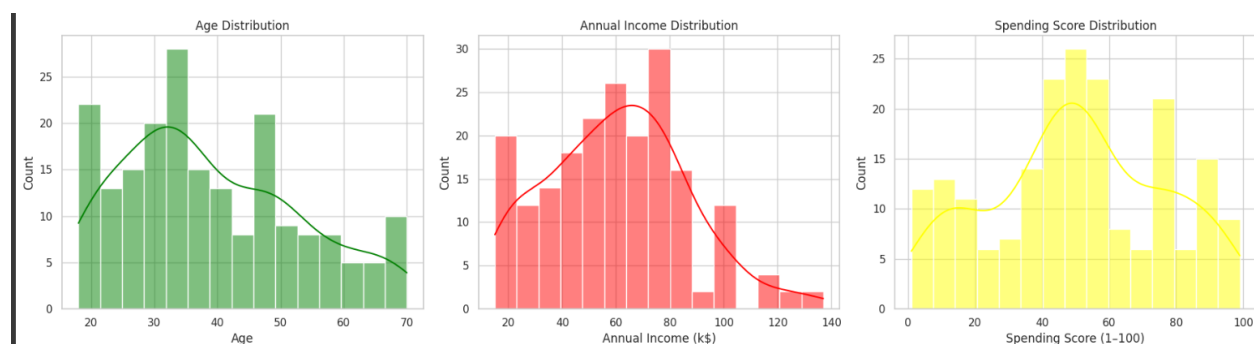

Figure 1: Distributions of Age, Annual Income, and Spending Score

- **Age:** The age distribution is right-skewed, with most customers aged between 25 and 40. Additionally, a significant number of customers fall within the 30–35 age range.

- **Annual Income (k$):** The annual income distribution appears approximately normal with a slight right skew. Most customers earn between 40k$ and 80k$ annually, with a peak around 70k$.

- **Spending Score (1–100):** The spending score is more uniformly distributed, with noticeable peaks around 40–60 and another around 75–80, suggesting the presence of distinct customer spending behavior groups.

3 Methodology

3.1 K-Means Objective Function

The K-Means algorithm aims to partition the dataset into k distinct clusters by minimizing the **within-cluster sum of squares (WCSS)**, which quantifies the total variance within each cluster. The objective function is defined as:

$$\text{WCSS} = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

Where:

- k is the number of clusters.

- C_j is the j-th cluster.

- $x_i \in C_j$ are the data points assigned to cluster C_j.

- μ_j is the centroid (mean) of cluster C_j.

The algorithm iteratively assigns each data point to the nearest centroid and then updates the centroids to be the mean of the assigned points. This process continues until convergence, typically when centroid positions no longer change significantly or after a set number of iterations.

3.2 Initialization Methods

Since K-Means is sensitive to the initial placement of centroids, different initialization methods can significantly impact both convergence speed and final clustering quality. We implemented and compared the following initialization strategies:

- **Random Initialization:** Centroids are randomly selected from the dataset.

- **K-Means++ Initialization:** This method aims to spread out the initial centroids more evenly. The first centroid is chosen randomly, and subsequent centroids are selected with a probability proportional to the squared distance from the nearest existing centroid.

- **Max-Distance Initialization:** This strategy selects the first centroid randomly and then iteratively chooses the next centroid as the data point that is farthest (in terms of Euclidean distance) from all previously selected centroids.

4 Results and Interpretation

4.1 Evaluation Methods

4.1.1 Elbow Method

The Elbow Method was used to identify the optimal number of clusters by plotting the WCSS against k.

(a) Random Initialization

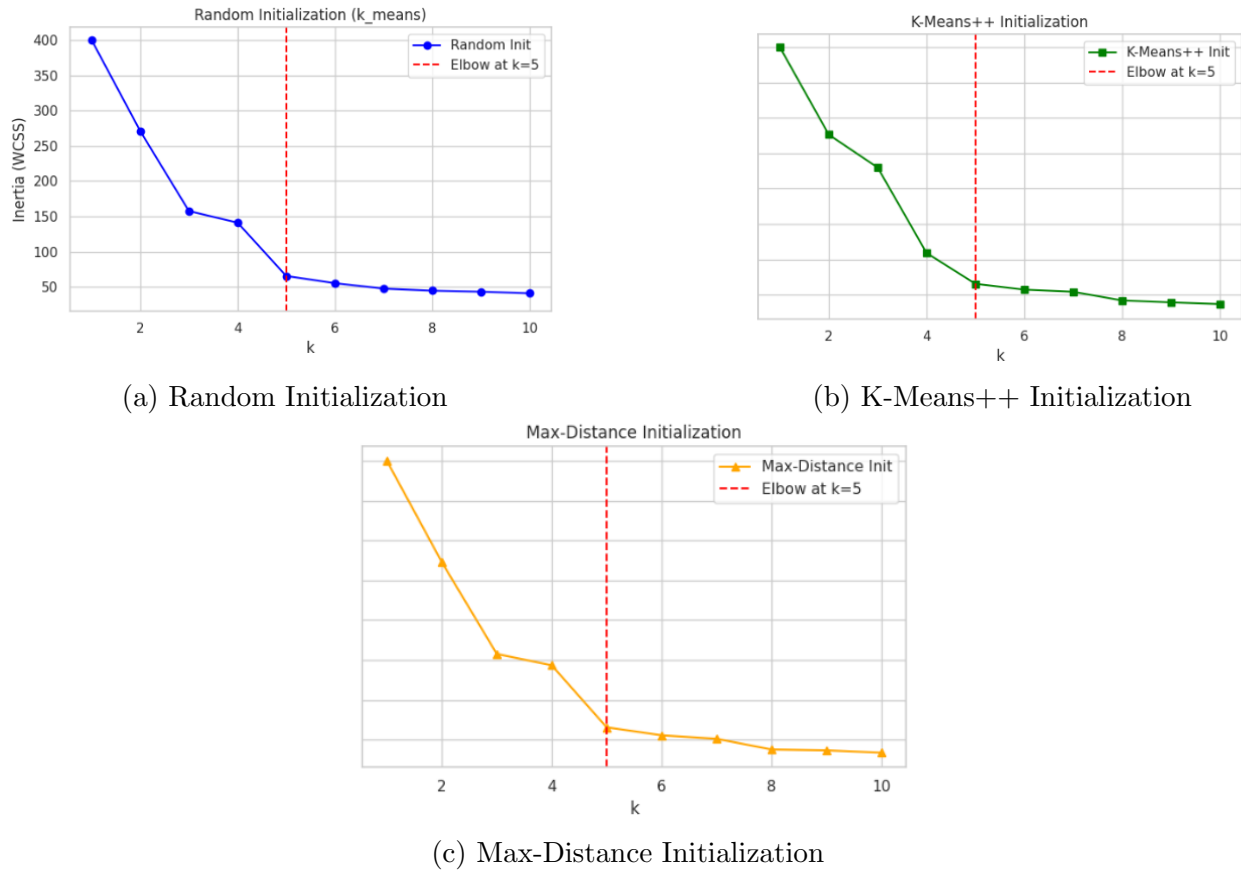(b) K-Means++ Initialization

(c) Max-Distance Initialization

Figure 2: Elbow Method for Different Initialization Strategies

As shown in Figure 2, the Elbow Method was applied to each initialization methods to determine the optimal number of clusters. For the Random Initialization method (Figure 2a), the elbow is observed at $k = 5$, suggesting five optimal clusters. The K-Means++ Initialization (Figure 2b) indicates an optimal $k = 5$, while the Max-Distance Initialization (Figure 2c) again suggests $k = 5$.

4.1.2 Silhouette Score

The silhouette score was used to assess clustering quality. Higher values indicate better-defined clusters.

Table 1: Silhouette Scores for Different Initialization Methods

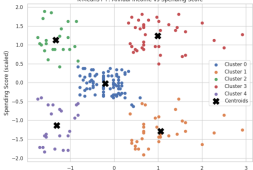| Initialization Strategy	Silhouette Score
K-Means (Random Init)	0.5539
K-Means++ Init	0.5547
Max-Distance Init	0.5547

As shown in Table 1, the K-Means++ and Max-Distance initialization methods yield slightly better clustering performance than Random Initialization, achieving a silhouette score of 0.5547 compared to 0.5539. This suggests that both K-Means++ and Max-Distance approaches lead to more coherent and well-separated clusters, likely due to better initial centroid placement.

4.2 Comparison of K-Means with Random Initialization against KMeans++

The overall clustering results demonstrated distinct separation among customer groups, reflecting varying spending behaviors and income levels across the different initialization methods.

4.2.1 Silhouette Scores Based on Annual Income vs Spending Score

Table 2: Silhouette Scores for Clustering Based on Annual Income vs Spending Score

Initialization Strategy	Silhouette Score	Visualization
K-Means	0.5539	
K-Means++	0.5547	

For the clustering based on Annual Income vs Spending Score, Table 2 presents the Silhouette Scores obtained using the K-Means and K-Means++ algorithms. The K-Means implementation from scratch achieves a Silhouette Score of 0.5539, while K-Means++ attains a

slightly higher score of 0.5547. These results indicate that both methods yield reasonably well-separated and compact clusters. However, the marginal improvement from K-Means++ suggests that its initialization strategy provides a slight advantage in forming more distinct and cohesive clusters, as also illustrated in the clustering visualizations provided alongside the scores.

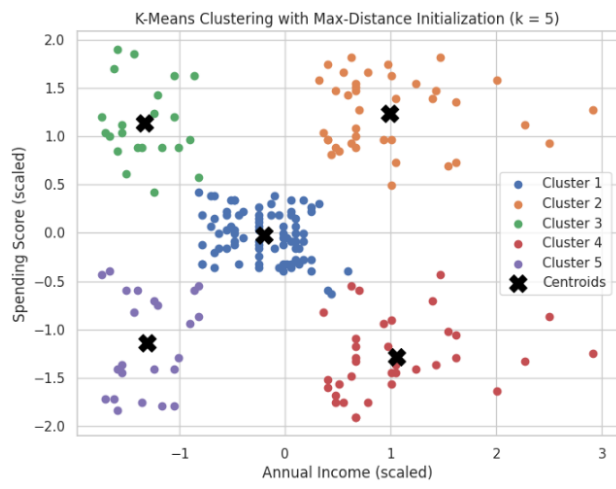### 4.2.2 Clustering Results for Different Initialization Methods

The K-Means Clustering algorithms with different initialization methods plots of Spending Score vs Annual Income from the Mall Customers dataset identifies distinct customer groups, such as low-income low spenders, high-income low spenders, low-income high spenders, and high-income high spenders each representing different shopping behaviors as shown in Figure 3. This segmentation helps businesses tailor marketing strategies and better understand customer value.
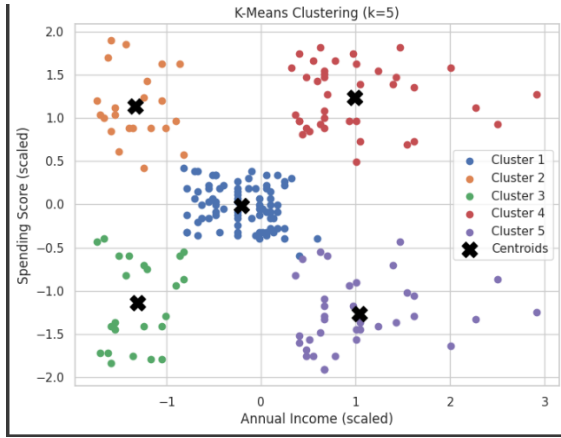
(a) K-Means (Random Init), $k = 5$
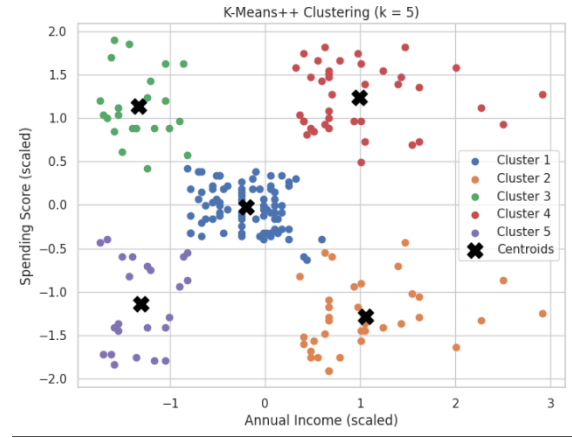
(b) K-Means++ Init, $k = 5$

(c) Max-Distance Init, $k = 5$

Figure 3: Customer Segmentation Results using Different Initialization Methods with Optimal k

Figure 3 shows the clustering results using three different centroid initialization methods in the K-Means algorithm, all of which yield an optimal number of clusters at $k = 5$. The Random Initialization method (Figure 3a) produces distinct customer segments based on annual income and spending score. The K-Means++ Initialization (Figure 3b) achieves well-separated clusters, demonstrating the effectiveness of smarter initialization. The Max-Distance Initialization method (Figure 3c) also results in five clearly defined clusters with strong separation.

5   Conclusion

This report demonstrated how K-Means clustering can effectively segment customers based on spending behavior and income. The K-Means++ initialization provided the most stable and high-quality clustering in this analysis.

Appendix

Access Notebook with Implementation Here

References

[1]   Kristina P. Sinaga and Miin-Shen Yang. "Unsupervised K-Means Clustering Algorithm". In: *IEEE Access* 8 (2020), pp. 80716–80727. DOI: 10.1109/ACCESS.2020.2988796.