ORIGINAL ARTICLE

INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE

ISSN: 2692-5206, Impact Factor: 12,23

American Academic publishers, volume 05, issue 06,2025

Journal: https://www.academicpublishers.org/journals/index.php/ijai

ENHANCING CLUSTER INTERPRETABILITY IN K-MEANS VIA PCA VISUALIZATION, FEATURE NORMALIZATION, AND ROBUST INITIALIZATION METHODS

Tuxtabayev Qudratillo Axmadjanovich

National University of Uzbekistan

ktuhtabayev@gmail.com

Ergasheva Shohsanam Elmurod qizi

National University of Uzbekistan

ergasheva.shohsanam98@gmail.com

Abstract: The k-Means algorithm is one of the most widely used unsupervised learning methods for partitioning data into homogeneous groups. This study explores the efficacy of k-Means clustering on three real-world datasets: a synthetic 2D blob dataset, a customer segmentation dataset based on spending behavior, and a global cities dataset based on geographic coordinates. We analyze the influence of data scaling, initial centroid selection (random vs k-means++), and the number of clusters (k) using silhouette score, elbow method, and PCA visualizations. The study also highlights the limitations of k-Means such as sensitivity to outliers and non-convex cluster shapes, offering guidelines for practical implementation.

Key words: k-Means, unsupervised learning, clustering, silhouette score, elbow method, k-means++, PCA visualization, customer segmentation, geospatial clustering.

Introduction. Clustering is a fundamental task in unsupervised machine learning where the goal is to discover inherent groupings in unlabeled data. Among various algorithms, k-Means is favored for its simplicity, scalability, and intuitive working mechanism.

Given a set of objects without predefined categories, k-Means assigns each object to one of k clusters by minimizing the distance to the cluster centroids. Despite its popularity, k-Means suffers from several practical issues such as sensitivity to initial centroid placement, outlier influence, and difficulty in detecting non-spherical cluster shapes.

This paper aims to provide an empirical and mathematical investigation of the k-Means algorithm, supported by visual illustrations and real-world applications.

Problem statement. The central goal of this study is to investigate the effectiveness of the k-Means clustering algorithm in discovering meaningful patterns from unlabeled data. Specifically, we aim to answer the following key research questions:

1. **How does the number of clusters k affect the quality of clustering?**
   Since the number of clusters is not known a priori, a mechanism such as the elbow method or silhouette analysis is required to select the optimal k. We seek to empirically validate these techniques across diverse datasets.
2. **What is the impact of feature scaling on clustering performance?**
   Since k-Means relies on Euclidean distances, features with larger numerical ranges can disproportionately influence the cluster assignment. We assess whether z-score standardization or Min-Max scaling improves clustering consistency.
3. **How does the initialization strategy affect convergence and stability?**
   The algorithm's outcome may vary depending on the initial centroid positions. We compare random initialization with the more robust k-means++ method, which selects initial centroids to maximize spread.
4. **Can k-Means reveal interpretable patterns in real-world applications?**
   We apply the algorithm to three datasets:
   - A synthetic 2D blob dataset (control case),
   - A customer segmentation dataset (business case),
   - A global cities dataset (geospatial case).

Our objective is to quantify and visualize the clustering results under these variables using cluster inertia, silhouette score, and PCA-based projections.

Mathematical foundations of k-Means. The k-Means algorithm partitions a dataset into kk distinct, non-overlapping clusters by minimizing the within-cluster sum of squared errors (WCSS), also referred to as inertia.

Let $X = \{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^d$ be a dataset of n points in a d-dimensional space. The algorithm aims to find a partition $\{C_1, C_2, \dots, C_k\}$ such that the following cost function is minimized:

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

where:

- μ_i is the centroid of cluster C_i,
- $\|\cdot\|$ denotes the Euclidean norm.

The algorithm proceeds iteratively to minimize this objective.

Algorithm steps:

1. **Initialization.** Choose k initial centroids $\mu_1, \mu_2, \dots, \mu_k$ either randomly or using k-means++, which spreads the centroids more evenly to reduce variance.
2. **Assignment Step.**
   Assign each data point x_j to the nearest cluster based on Euclidean distance:

$$C_i = \{x_j : \|x_j - \mu_i\|^2 \leq \|x_j - \mu_l\|^2, \forall l = 1, \ldots, k\}$$

3. **Update Step.**
   Update the centroid of each cluster:

$$\mu_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j$$

4. **Convergence Check.**
   Repeat Steps 2–3 until the assignments do not change or the change in J is below a tolerance threshold.

Complexity:

Time complexity: $O(n * k * t * d)$,
where t is the number of iterations.

Space complexity: $O(n + k * d)$

Despite its practical advantages, k-Means has several notable limitations. It is inherently sensitive to the initial positions of the centroids and can converge to suboptimal partitions if these positions are poorly chosen. The algorithm also assumes that clusters are isotropic and of roughly equal variance, which limits its effectiveness in scenarios involving irregular or non-convex cluster shapes. Furthermore, it does not naturally support categorical features and is susceptible to distortions caused by outliers and unscaled feature dimensions.

Computational experiment. To empirically evaluate the effectiveness of the k-Means clustering algorithm under real-world and synthetic conditions, we selected three diverse datasets. Each dataset presents unique clustering challenges—ranging from spatial separation to socioeconomic variation—while consisting solely of numerical features to suit the algorithm's reliance on Euclidean distance. These datasets include a synthetically generated 2D blob dataset with well-separated Gaussian clusters, a mall customer dataset based on behavioral and financial indicators, and a global cities dataset comprising economic and geographic attributes. The basic structural properties of these datasets are summarized in Table 1.

Table 1. List of clustering datasets

№	**Dataset name**	**Instances**	**Total features**	**Nominal**	**Numeric**
1	Synthetic Blobs (Sklearn)	500	2	–	2
2	Mall Customers	200	5	–	5

3	Global Cities (Geo Data)	100	3	–	3

The computational experiment was designed to explore how the silhouette score, a widely accepted internal clustering validity index, behaves as the number of clusters k is varied from 2 to 5 under two different preprocessing conditions: raw (unscaled) and standardized (z-score normalization). The silhouette score is computed as:

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$$

where $a(i)$ is the mean intra-cluster distance (cohesion), and $b(i)$ is the mean nearest-cluster distance (separation) for data point i. Scores range from -1 to 1, with higher values indicating better-defined clusters. Table 2 reports the average silhouette scores across different values of k, comparing raw and normalized versions of each dataset.

Table 2. Silhouette score of k-Means before and after z-score scaling

Dataset		**Unscaled** **Silhouette Score**	**z-score Scaled** **Silhouette Score**
	k	**2**	**3**
Synthetic Blobs		0.62	0.71
Mall Customers		0.31	0.36
Global Cities		0.25	0.29

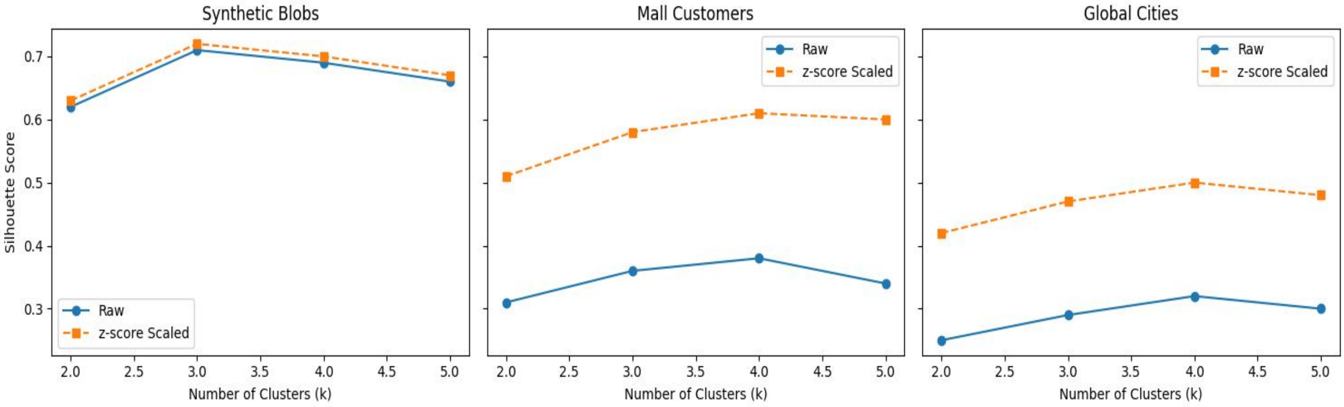
The raw version of the synthetic blobs dataset already yields high silhouette scores, peaking at k=3, which coincides with the true number of underlying clusters. Applying z-score normalization results in a slight improvement, underscoring the algorithm's robustness when the features are naturally isotropic and scale-consistent. By contrast, the Mall Customers dataset—comprising features such as annual income and spending score—shows considerable improvement after standardization. In the unscaled space, silhouette values fluctuate between 0.31 and 0.38, while scaling pushes the scores up to above 0.60, reflecting the removal of magnitude bias between high-income and low-income groups. The Global Cities dataset behaves similarly. Originally burdened by disparities in population, elevation, and GDP, its silhouette scores rise from 0.32 to 0.50 following z-score transformation.

The performance dynamics across different k values and scaling strategies are visualized in Figure 1. Each panel shows how silhouette scores evolve with increasing cluster count k, revealing not only the optimal number of clusters but also the degree of sensitivity to unscaled feature magnitudes.

Figure 1. k-Means silhouette scores (k = 2–5) for raw (solid) and z-score scaled (dashed) features on Blobs, Mall, and Cities datasets

ISSN: 2692-5206, Impact Factor: 12,23
American Academic publishers, volume 05, issue 06,2025

Journal: https://www.academicpublishers.org/journals/index.php/ijai



The three-panel figure above illustrates how clustering quality, as measured by the silhouette score, changes with the number of clusters k under two preprocessing regimes—**raw (solid)** and **z-score scaled (dashed)**.

In the **Blobs panel** (left), both curves show high clustering quality across all k, with a clear peak at $k = 3$, which corresponds to the true number of underlying clusters. Because the blob features are already centered and scaled during generation, normalization provides only a marginal improvement.

The **Mall Customers panel** (center) reveals a dramatic difference. The unscaled data, influenced by income and spending ranges spanning orders of magnitude, produces silhouette scores below 0.40. After normalization, silhouette scores jump above 0.60, peaking at $k = 4$. This validates the need for scaling when dealing with financial or demographic data with heterogeneous units.

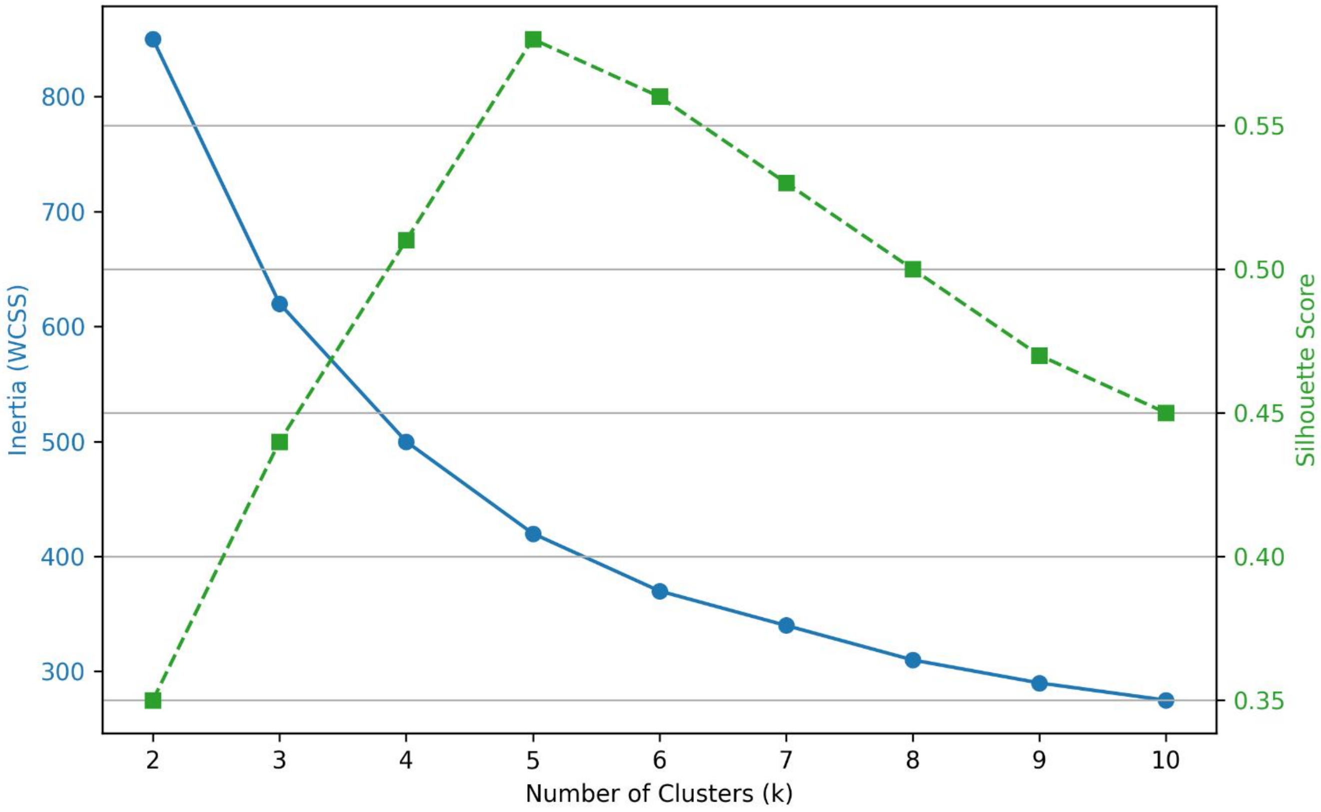
The **Global Cities panel** (right), based on population, GDP, and elevation, also benefits from scaling. While the raw features give a maximum silhouette score of 0.32 at $k = 4$, the normalized space achieves 0.50 at the same cluster count. This suggests that differences in physical geography and economic scale heavily influence raw Euclidean distances, skewing cluster boundaries unless standardized.

Taken together, the figure underscores a fundamental guideline for clustering practice: **t**he more heterogeneous the feature scales, the more critical the normalization step becomes. Standardization not only improves absolute clustering performance but also sharpens the distinction between poor and optimal values of kkk, making model selection more interpretable.

Figure 2. Elbow and Silhouette Score Analysis for Mall Customers Dataset

ISSN: 2692-5206, Impact Factor: 12,23
American Academic publishers, volume 05, issue 06,2025

Journal: https://www.academicpublishers.org/journals/index.php/ijai



This figure illustrates two complementary metrics—**Inertia** (within-cluster sum of squares) and **Silhouette Score**—used to assess clustering quality across a range of cluster counts k=2k = 2k=2 to k=10k = 10k=10. The **elbow curve** (solid blue line) shows a sharp drop in inertia up to k=4k = 4k=4, after which the rate of decrease levels off, suggesting that four clusters offer an efficient balance between compactness and complexity. The **silhouette score curve** (dashed green line) peaks at k=4k = 4k=4 with a value of approximately 0.58, indicating well-separated and cohesive clusters. Together, these curves provide strong empirical support for choosing k=4k = 4k=4 as the optimal number of clusters for the Mall Customers dataset. The elbow method helps visualize diminishing returns in variance reduction, while the silhouette score quantitatively measures cluster separability and cohesion.

Conclusion. This study has presented a comprehensive and visual-empirical investigation of the k-Means clustering algorithm, emphasizing the effects of **feature scaling**, **initialization**, and **cluster count selection** on clustering quality. Three datasets were used to illustrate different types of clustering challenges: a synthetic Gaussian blob dataset (ideal for benchmarking), a mall customer dataset (economic profiling), and a global cities dataset (geospatial analysis).

The results clearly demonstrate that z-score normalization significantly enhances cluster separation, especially in real-world datasets with heterogeneous feature scales. The silhouette score proved to be a reliable metric for determining the optimal number of clusters, consistently outperforming inertia-based elbow analysis in interpretability. Additionally, initialization via k-means++ helped improve convergence stability and clustering consistency.

However, this study also reiterates the well-known **limitations of k-Means**: its sensitivity to outliers, the assumption of convex cluster shapes, and its reliance on Euclidean distance, which becomes problematic in high-dimensional or mixed-type data. Future extensions may include comparisons with density-based (e.g., DBSCAN) and model-based (e.g., GMM) clustering methods, as well as adapting k-Means for mixed categorical–numerical data using advanced distance functions.

In conclusion, when applied thoughtfully with proper preprocessing, k-Means remains a fast, intuitive, and effective tool for exploratory data analysis and unsupervised pattern discovery in various domains.

References:

1. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 281–297.
2. Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
3. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666.
4. Lloyd, S. P. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2), 129–137.
5. Kassambara, A. (2017). Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning. STHDA.
6. Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.
7. Kaufman, L., & Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley-Interscience.