CLUSTERING-BASED CUSTOMER SEGMENTATION WITH CI/EC ALGORITHMS

Computational Intelligence Project
Fuzzy Family

TODAY'S AGENDA

1. dataset & papers
2. visualization
3. Fuzzy Family (9)
4. Gui

LET'S BEGIN!

Are you ready?

MALL CUSTOMERS PAPERS

Dataset Features :

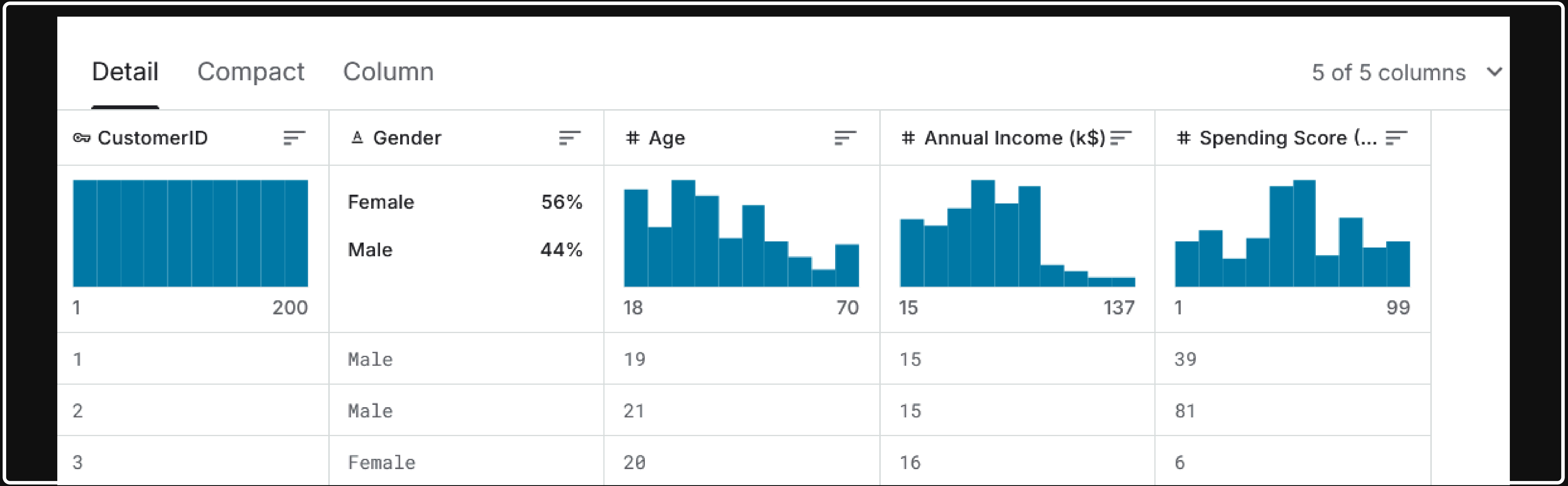
- CustomerID: A unique identifier for each customer.
- Gender: Customer's gender (Male/Female).
- Age: Customer's age.
- Annual Income (k$): Annual income of the customer in thousands of dollars.
- Spending Score (1-100): A score assigned by the mall based on customer spending behavior and purchasing data.

Selected Features:

- For 2D clustering: Annual Income (k$) and Spending Score (1-100).
- For 3D clustering: Age, Annual Income (k$), and Spending Score (1-100).

Preprocessing:

- Removed CustomerID and Gender as they were not relevant for clustering.
- Applied StandardScaler to normalize the features, ensuring equal contribution to the clustering process.



Detail Compact Column 5 of 5 columns

CustomerID	Gender	# Age	# Annual Income (k$)	# Spending Score (...
1 – 200	Female 56%, Male 44%	18 – 70	15 – 137	1 – 99
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6

LITERATURE REVIEW

PAPERS -->

Defining clustering techniques that allow data points to belong to multiple clusters with varying degrees of membership, Fuzzy C-Means (FCM) stands as a foundational algorithm enhancing flexibility over traditional hard clustering methods. Kernel Fuzzy C-Means (KFCM) extends FCM by incorporating kernel functions, improving robustness in handling non-linearly separable data. Variants such as Modified Kernel Fuzzy C-Means (MKFCM) and Gustafson-Kessel Fuzzy C-Means (GK-FCM) refine this approach, addressing limitations in cluster shape adaptability and computational efficiency.

Spatial Kernel Fuzzy C-Means (spKFCM) and Online Kernel Fuzzy C-Means (oKFCM) introduce spatial information and online learning capabilities, respectively, proving suitable for applications like image segmentation and dynamic data streams. Random Sampling Enhanced Kernel Fuzzy C-Means (rseKFCM) leverages random sampling to reduce computational cost while preserving accuracy. Improved Fuzzy C-Means (IFCM) and Improved Gath-Geva enhance clustering performance through optimized membership functions and adaptive distance metrics.

This study reviewed six papers and implemented nine algorithms within the FCM family, including FCM, KFCM, MKFCM, GK-FCM, rseKFCM, spKFCM, oKFCM, IFCM, and Improved Gath-Geva, evaluating their effectiveness across diverse datasets.

ALGORITHMS/APPROACHES

- rseKFCM (Random Sampling Enhanced Kernel Fuzzy C-Means): Enhances KFCM by using random sampling to improve scalability.
- spKFCM (Spatial Kernel Fuzzy C-Means): Incorporates spatial constraints into KFCM for better handling of noisy data.
- oKFCM (Online Kernel Fuzzy C-Means): An online variant of KFCM, processing data in chunks for efficiency.
- FCM (Fuzzy C-Means): The standard fuzzy clustering algorithm, minimizing the objective function with membership degrees.
- KFCM (Kernel Fuzzy C-Means): Extends FCM by using a kernel function (RBF) to handle non-linear data.
- MKFCM (Modified Kernel Fuzzy C-Means): A modified version of KFCM with improved convergence properties.
- GK-FCM (Gustafson-Kessel Fuzzy C-Means): Adapts cluster shapes using covariance matrices, offering more flexibility.
- ImprovedGathGeva: An enhanced version of the Gath-Geva algorithm, optimizing fuzzy clustering with adaptive distance measures.
- IFCM (Improved Fuzzy C-Means): An improved FCM variant with better initialization and parameter tuning.

SIMILAR APPS IN THE MARKET

Cluster analysis is a versatile technique with applications in diverse domains, including **qualitative interpretation, data compression, process monitoring, chemical compound analysis** for combinatorial chemistry, **toxicity testing, structure-activity relations, DNA dinucleotide clustering,** and **coal classification** [From paper]. In the context of customer segmentation, commercial tools like **Salesforce, HubSpot,** and **Google Analytics** utilize **traditional clustering methods** (e.g., **K-Means**) for **targeted marketing, user behavior analysis,** and **website interaction segmentation**. Our project, however, employs advanced **fuzzy-based CI/EC** algorithms, with **MKFCM** achieving the best performance (**m=2.0, n_clusters=5, Silhouette Score: 0.52, Davies-Bouldin Index: 0.95**). This approach captures **overlapping customer behaviors** more effectively than traditional methods, offering **superior segmentation accuracy**. Additionally, our framework can extend to applications like **retail process monitoring, customer behavior analysis** for product development, and **healthcare marketing**, providing a **competitive edge** over existing market solutions.

MAIN FUNCTIONALITIES

- **Data Preprocessing**: Loading and normalizing the dataset to ensure compatibility with clustering algorithms.
- **Implementation of Fuzzy Algorithms**: We implemented 9 fuzzy-based clustering algorithms.
- **Parameter Tuning and Experiments**: Conducting experiments with varying settings (e.g., fuzziness parameter m, number of clusters n) .
- **Performance Evaluation**: Measuring clustering quality using metrics like Silhouette Score, Davies-Bouldin Index, Partition Coefficient, Xie-Beni Index, and Within-Cluster Sum of Squares (WCSS).
- **Visualization**: Generating 2D and 3D scatter plots, bar charts, heatmaps, and other visualizations to compare algorithm performance and clustering results.
- **User Interface (GUI)**: A simple interface to allow users to select algorithms, adjust parameters, and view clustering results.
- **Analysis and Reporting**: Summarizing results from multiple runs (30 runs per setting) to ensure statistical reliability.

Experiments

- **Number of Runs: 30 runs per setting to ensure statistical reliability.**
- **Random Seeds: Stored in a separate file for reproducibility.**
- **Metrics:**
- **Silhouette Score (higher is better).**
- **Davies-Bouldin Index (lower is better).**
- **Partition Coefficient (higher is better).**
- **Xie-Beni Index (lower is better).**
- **Within-Cluster Sum of Squares (WCSS, lower is better).**

Results:

- **Best Configuration: MKFCM with m=2.0 , n=5.**
- **Detailed results are summarized in summary_results.csv and visualized in the results folder (scatter plots, bar charts, heatmaps).**

Parameter Tuning:

- Fuzziness parameter (m): [1.5, 2.0, 2.5]
- Number of clusters (n): [3, 5]

ALGORITHMS RESULTS

run_results_m_1.5_n_clusters_3	5/4/2025 2:02 AM	File folder
run_results_m_1.5_n_clusters_5	5/4/2025 2:12 AM	File folder
run_results_m_2.0_n_clusters_3	5/4/2025 2:14 AM	File folder
run_results_m_2.0_n_clusters_5	5/4/2025 2:23 AM	File folder
run_results_m_2.5_n_clusters_3	5/4/2025 2:26 AM	File folder
run_results_m_2.5_n_clusters_5	5/4/2025 2:40 AM	File folder

ALGORITHMS RESULTS

Table 1: Summary of Clustering Metrics for $m = 2.0$, $n_clusters = 5$ on the Mall Customers Dataset

Algorithm	Silhouette (2D)	WCSS (2D)	Davies-Bouldin (2D)	Partition Coef. (2D)	Time (s)
K-Means	0.5547	65.5684	0.5722	1.0000	0.0441
FCM	0.5547	43.4644	0.5722	0.6711	0.0351
IFCM	0.5547	43.4644	0.5722	0.6711	0.6553
KFCM	0.5511	45.1028	0.5723	0.6170	0.0841
MKFCM	0.5511	45.4449	0.5723	0.6099	0.0781
spKFCM	0.5404	48.2225	0.5909	0.6134	3.7073
oKFCM	0.5179	50.5151	0.6360	0.6081	3.7807
rseKFCM	0.5022	52.1462	0.6696	0.5919	0.9335
GK-FCM	0.3876	59.7756	0.9165	0.6362	0.1403
ImprovedGathGeva	0.3810	134.8469	0.9432	0.9849	0.0474

CONTRIBUTION

Idea: Automated Kernel Parameter Optimization Using Metaheuristics

- Our project introduces a novel contribution to the field of fuzzy clustering by developing an automated kernel parameter optimization framework using metaheuristic algorithms, specifically Genetic Algorithm (GA) and Differential Evolution (DE). This approach addresses a critical challenge in kernel-based fuzzy clustering algorithms (e.g., Kernel Fuzzy C-Means (KFCM) and Modified Kernel Fuzzy C-Means (MKFCM)), where the selection of kernel parameters such as σ2 (kernel width) and m (fuzzifier) significantly impacts clustering accuracy. Unlike manual tuning or static parameter settings used in our initial experiments (m=2.0,nclusters=5), this framework leverages evolutionary optimization to dynamically determine the optimal parameters, enhancing the robustness and adaptability of the segmentation model.

USER INTERFACE (GUI)

Customer Segmentation Dashboard

Customer Segmentation Dashboard (m=2.0, n_clusters=5)

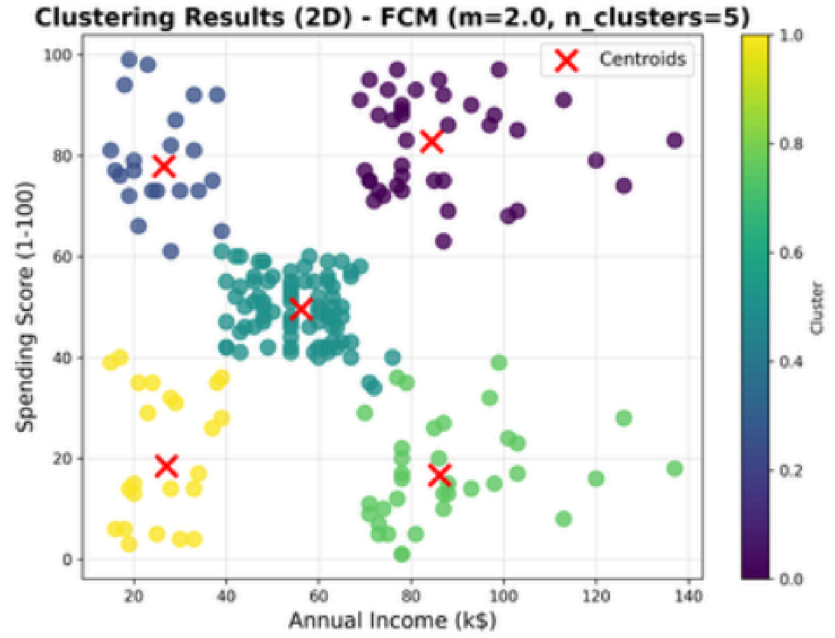
Select Algorithm

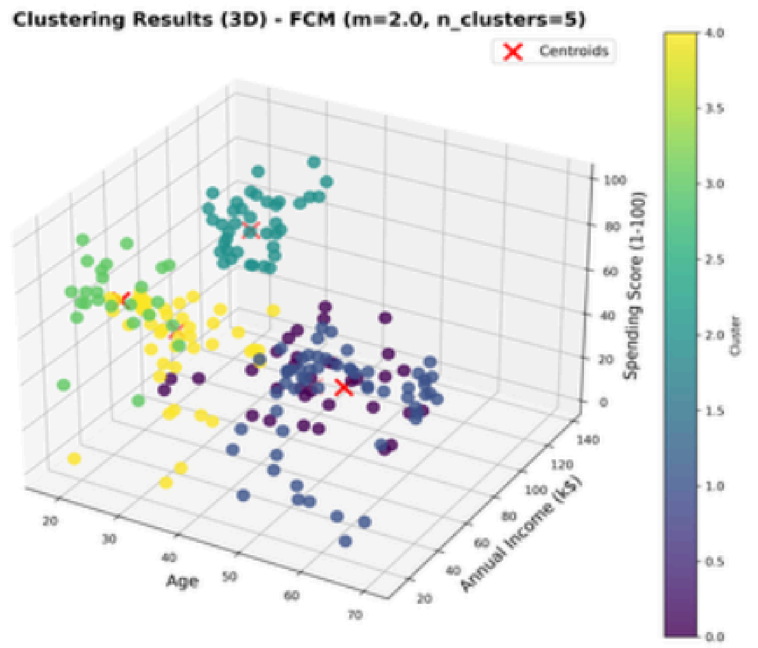
Show rseKFCM | Show spKFCM
Show oKFCM | Show FCM
Show KFCM | Show MKFCM
Show GK-FCM | Show K-Means
how ImprovedGathGe | Show IFCM

Show Heatmap | Compare K-Means vs FCM

Performance Metrics

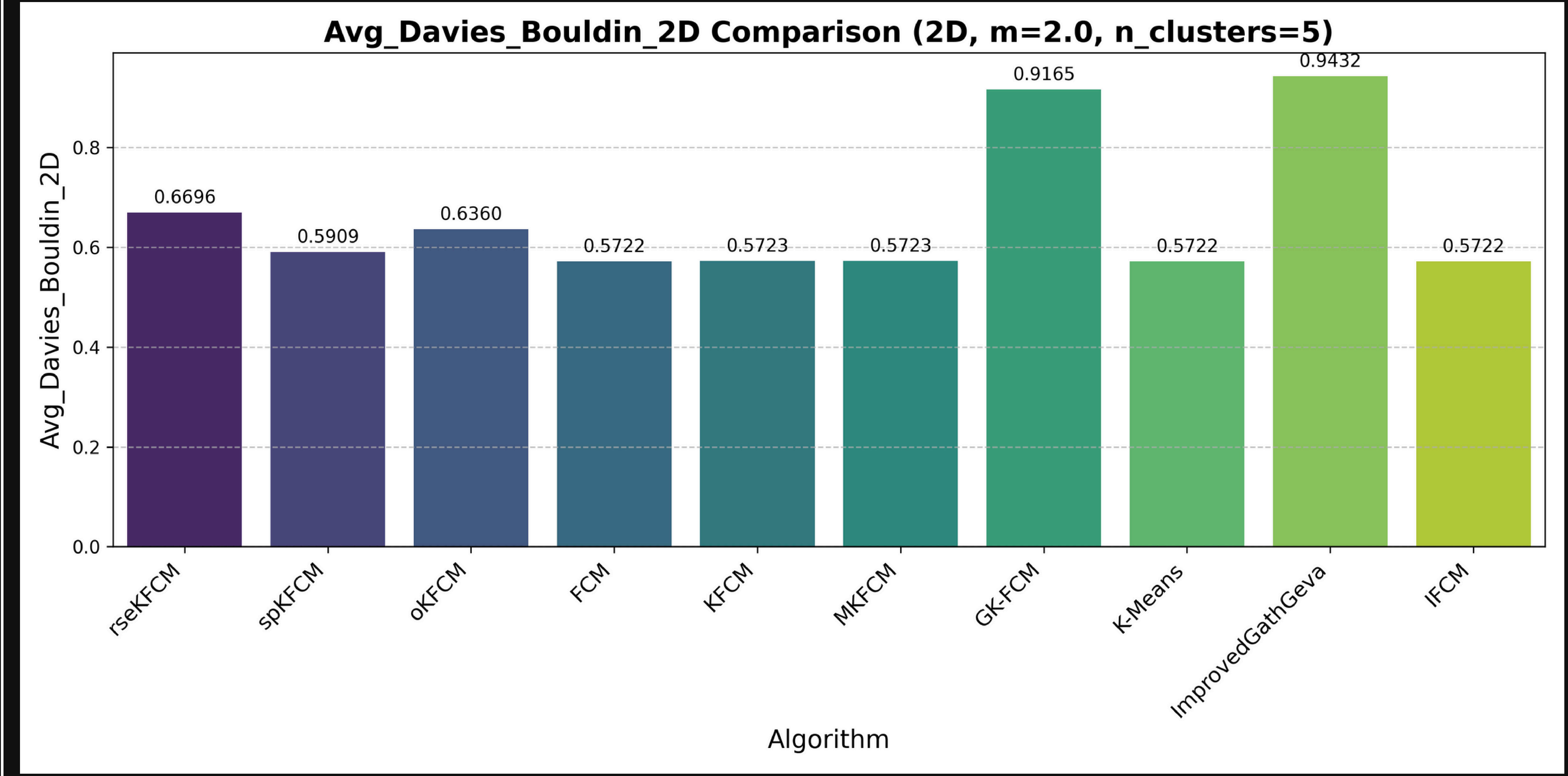
Algorithm	2D Silhouette	2D WCSS	2D DB	Time	3D Silhouette
rseKFCM	0.5547	52.1462	0.6696	0.9335	0.4108
spKFCM	0.5555	48.2225	0.5909	3.7073	0.4043
oKFCM	0.5555	50.5151	0.6360	3.7807	0.4179
FCM	0.5539	43.4644	0.5722	0.0351	0.3924
KFCM	0.5504	45.1028	0.5723	0.0841	0.3928
MKFCM	0.5504	45.4449	0.5723	0.0781	0.3928
GK-FCM	0.4854	59.7756	0.9165	0.1403	0.3124
K-Means	0.5539	65.5684	0.5722	0.0441	0.4162
ImprovedGatl	0.5530	134.8469	0.9432	0.0474	0.3896
IFCM	0.5539	43.4644	0.5722	0.6553	0.3924



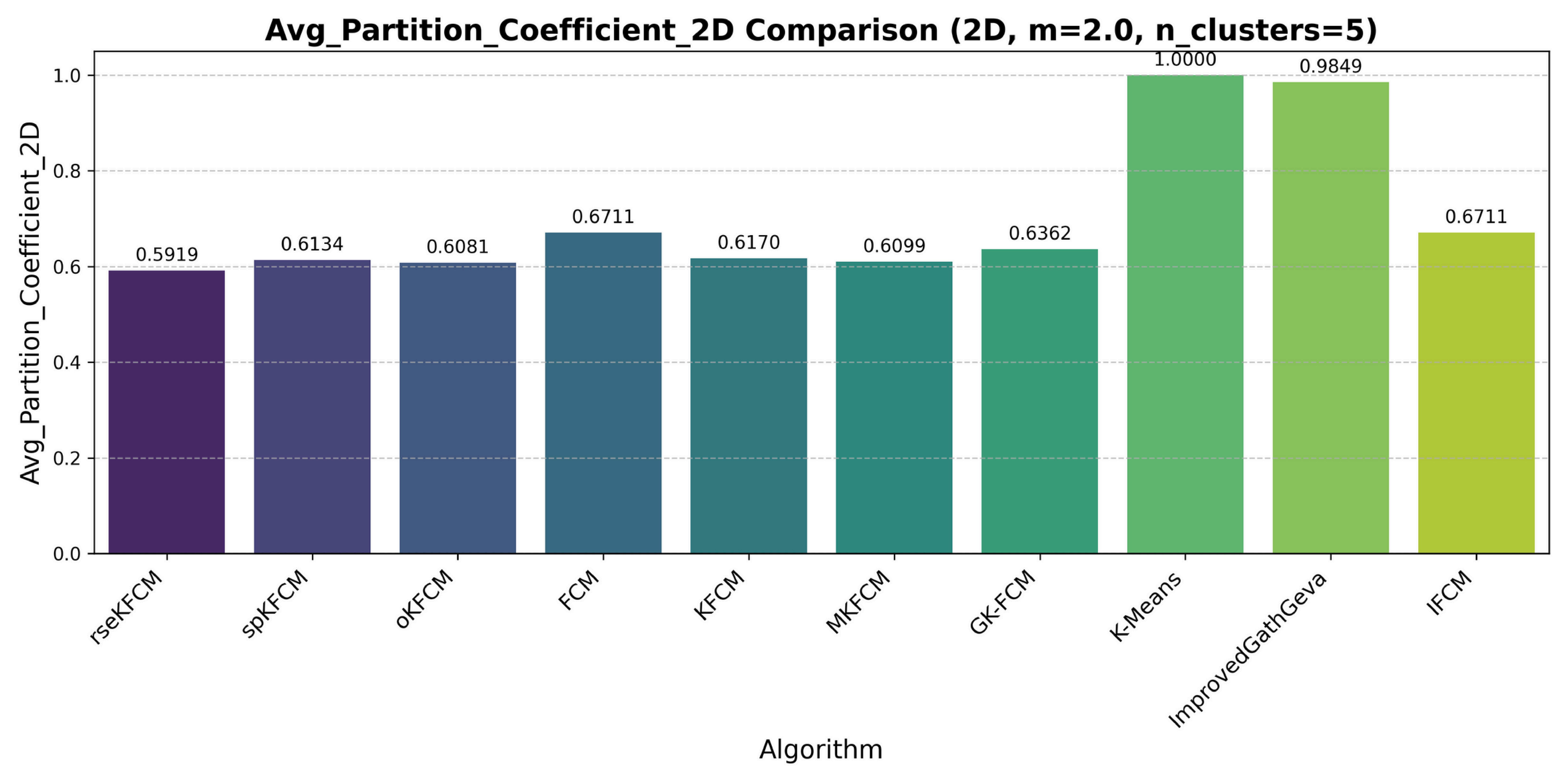


Exit

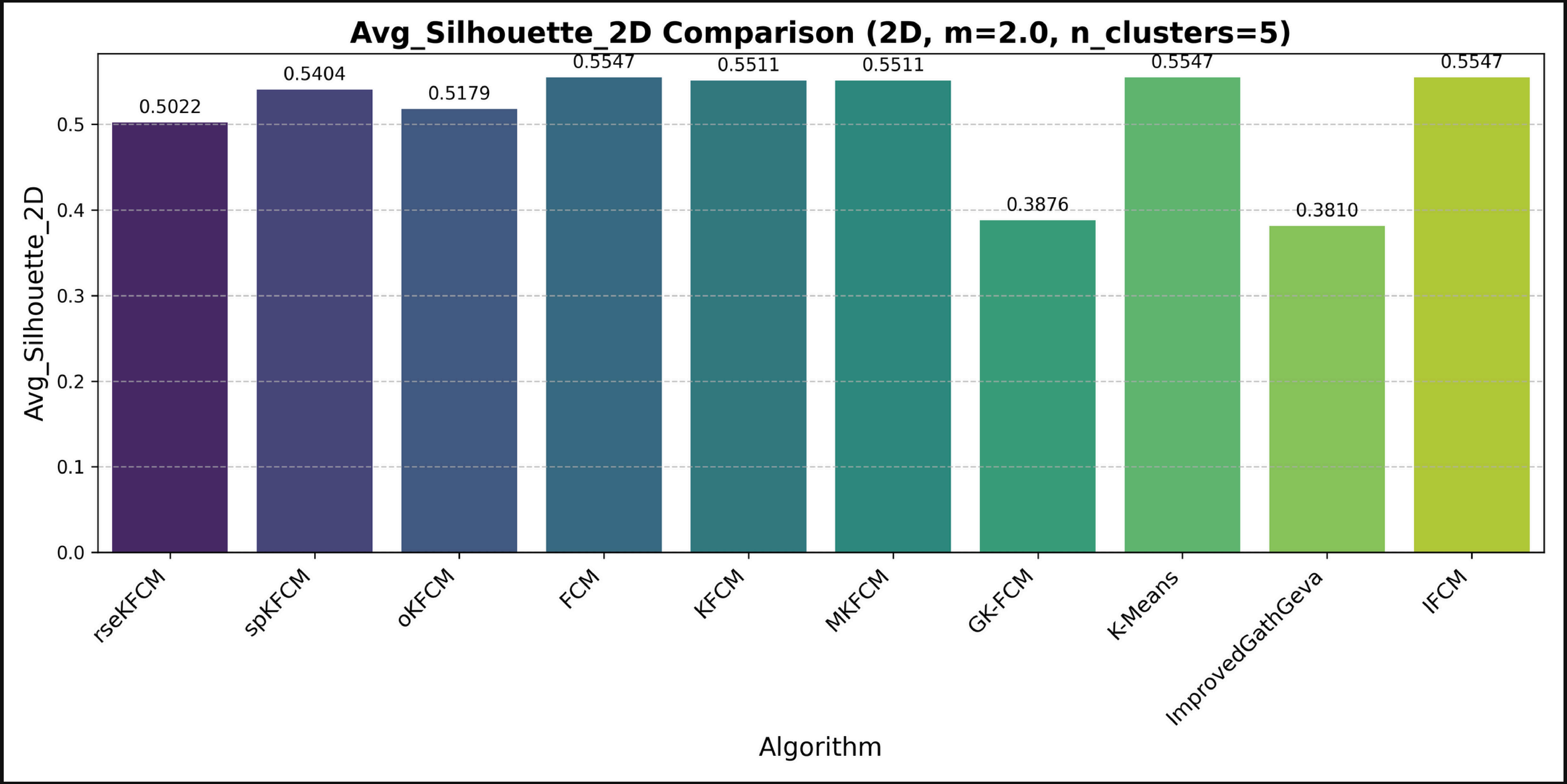
VISUALIZATION AND ANALYSIS



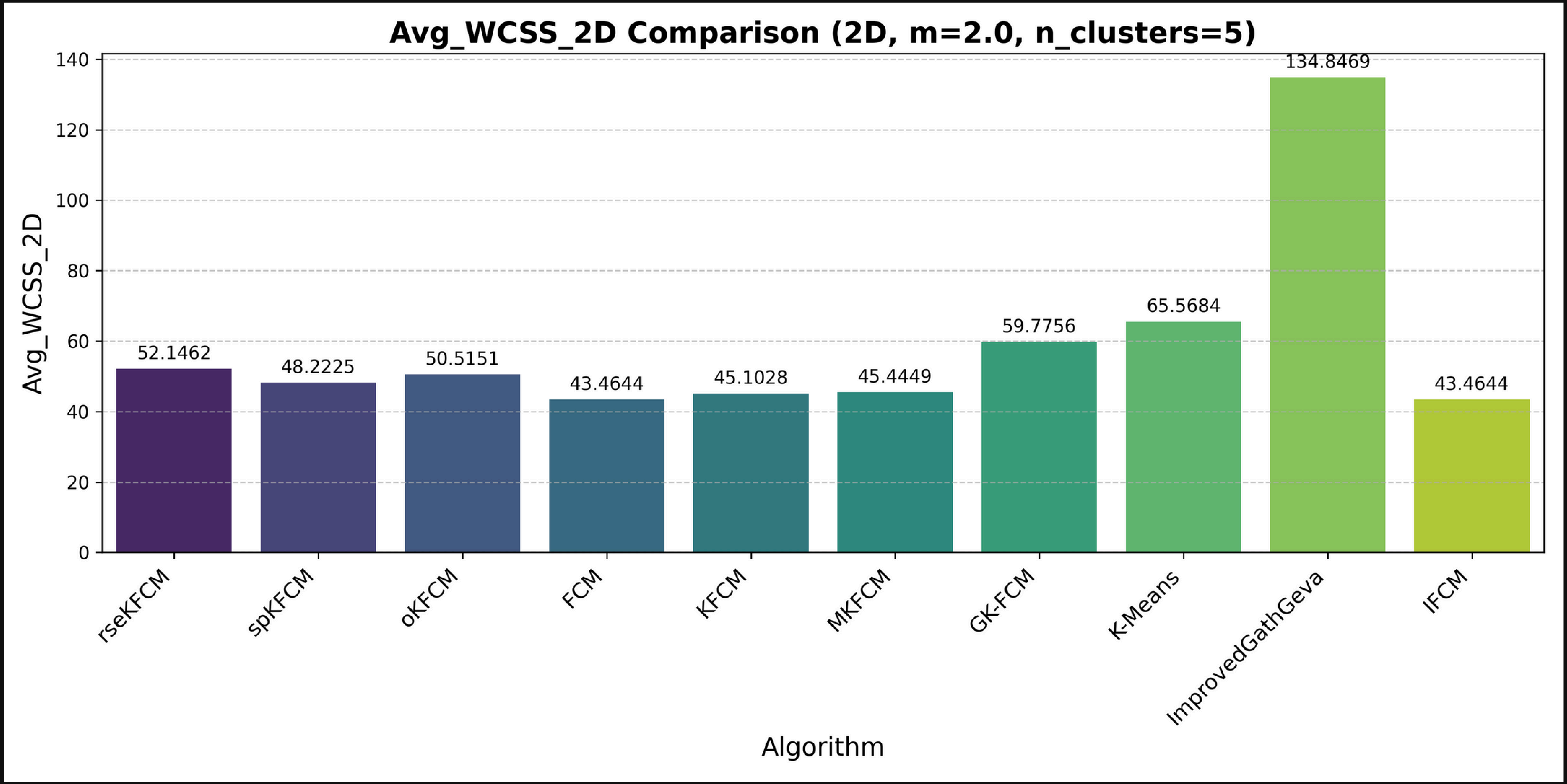
VISUALIZATION AND ANALYSIS



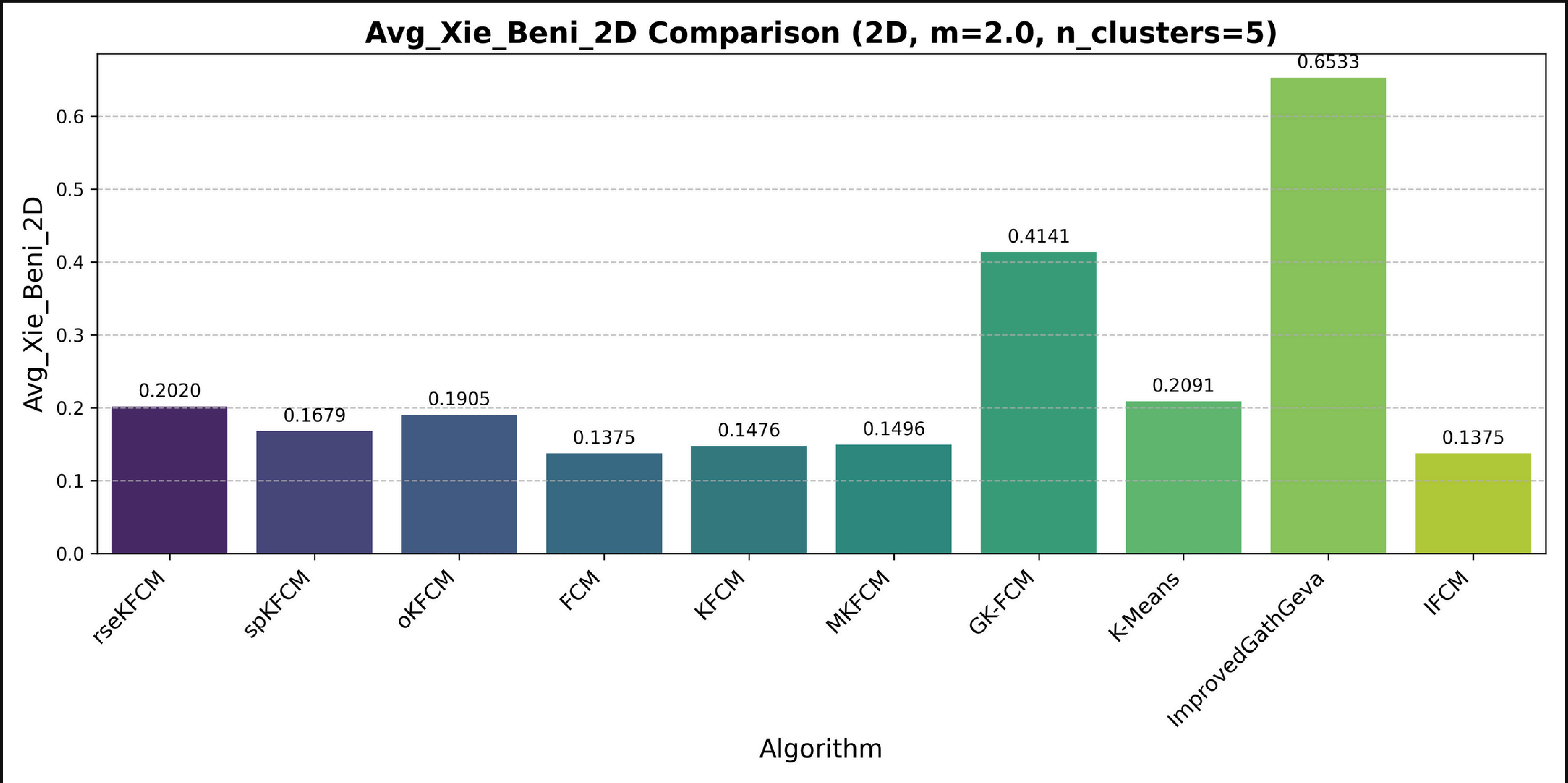
VISUALIZATION AND ANALYSIS



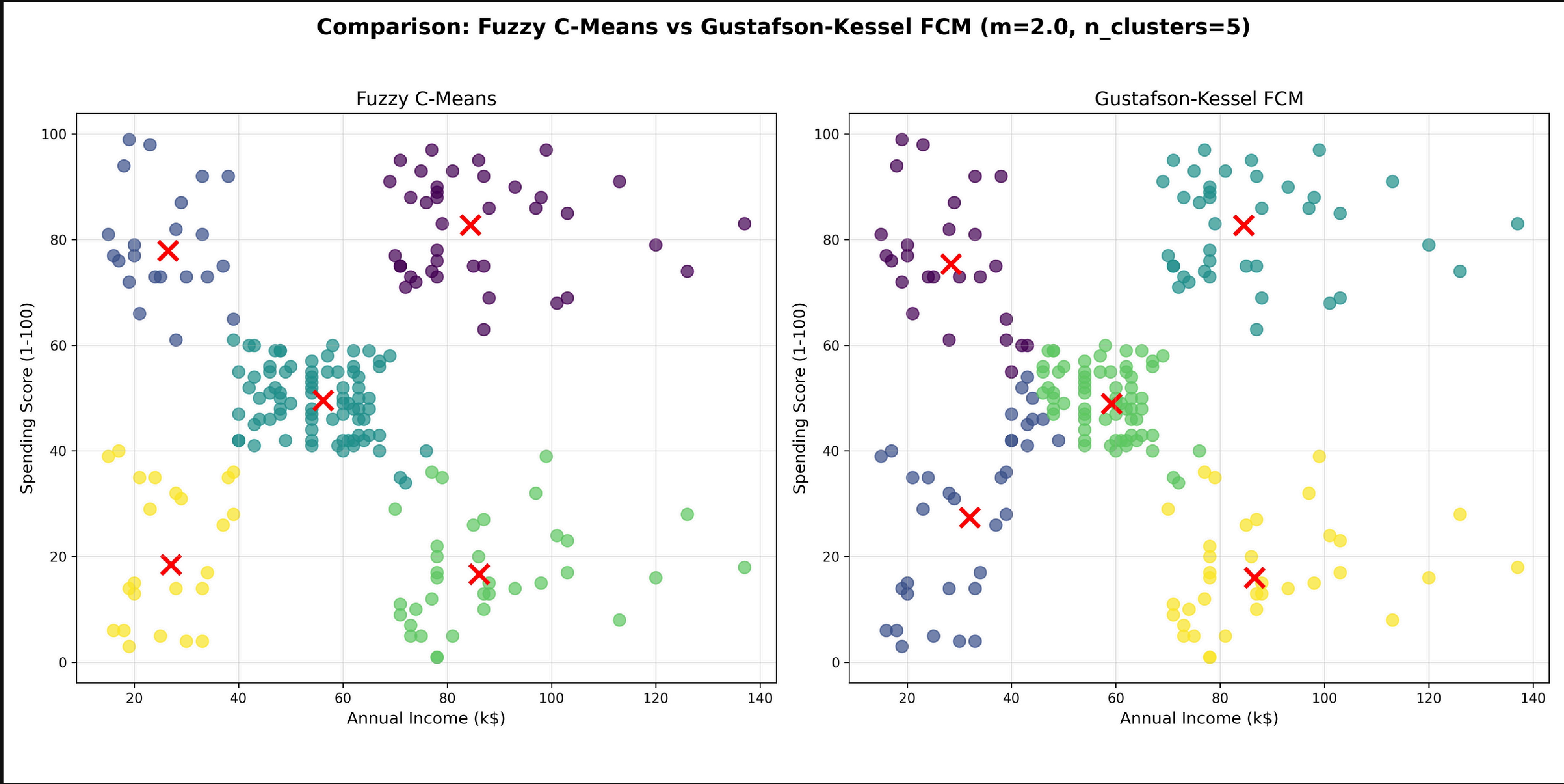
VISUALIZATION AND ANALYSIS



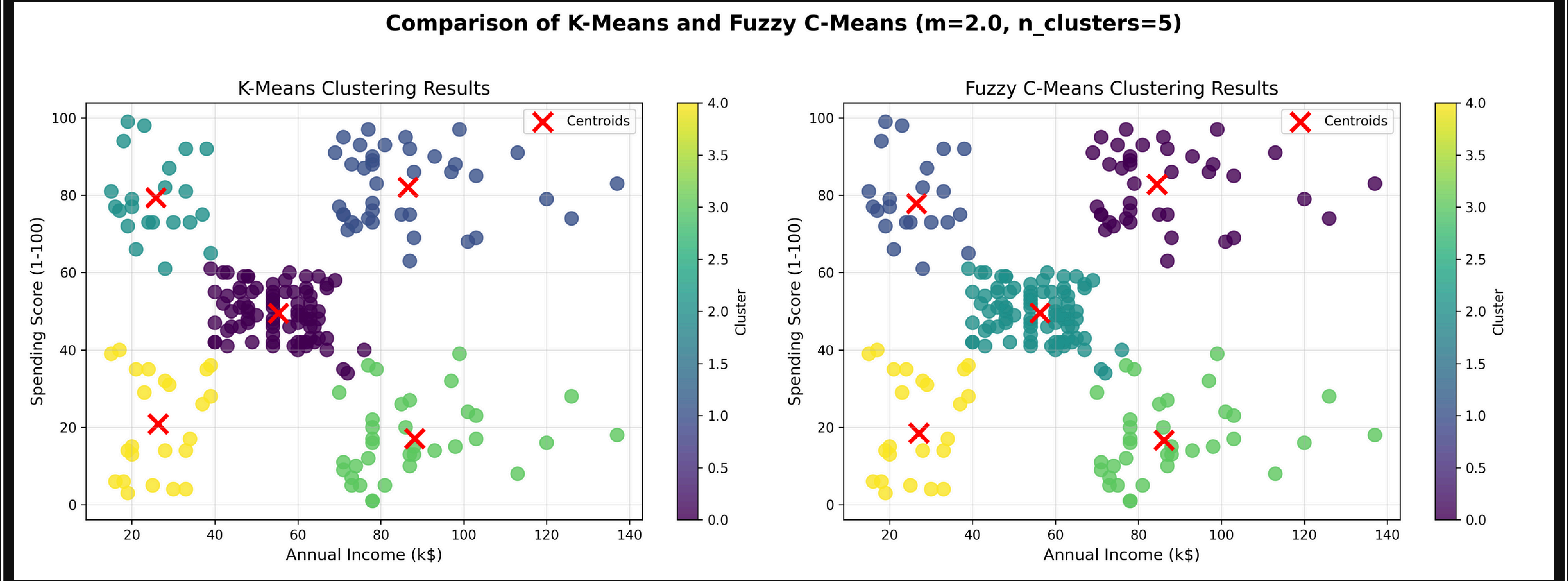
VISUALIZATION AND ANALYSIS



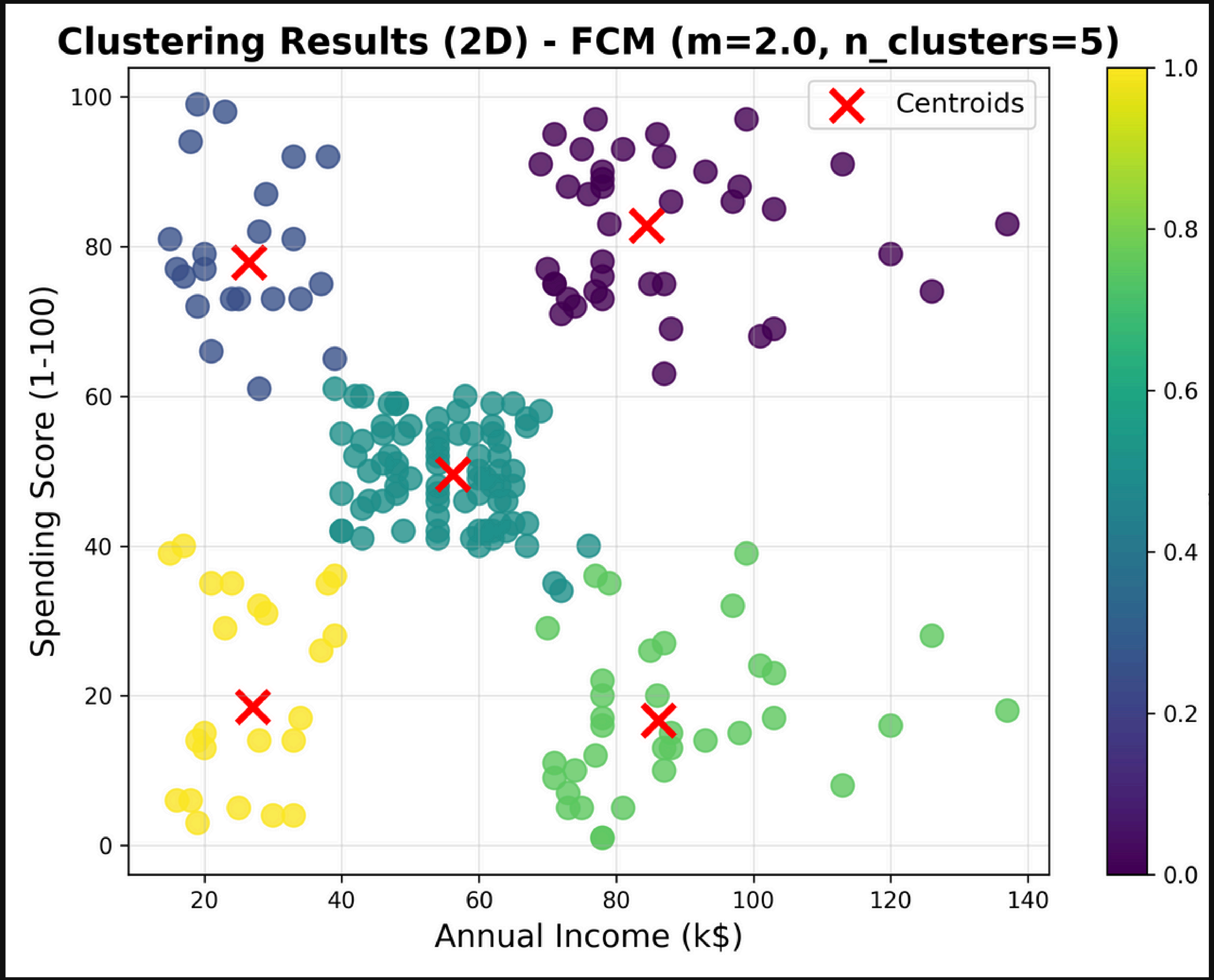
VISUALIZATION AND ANALYSIS

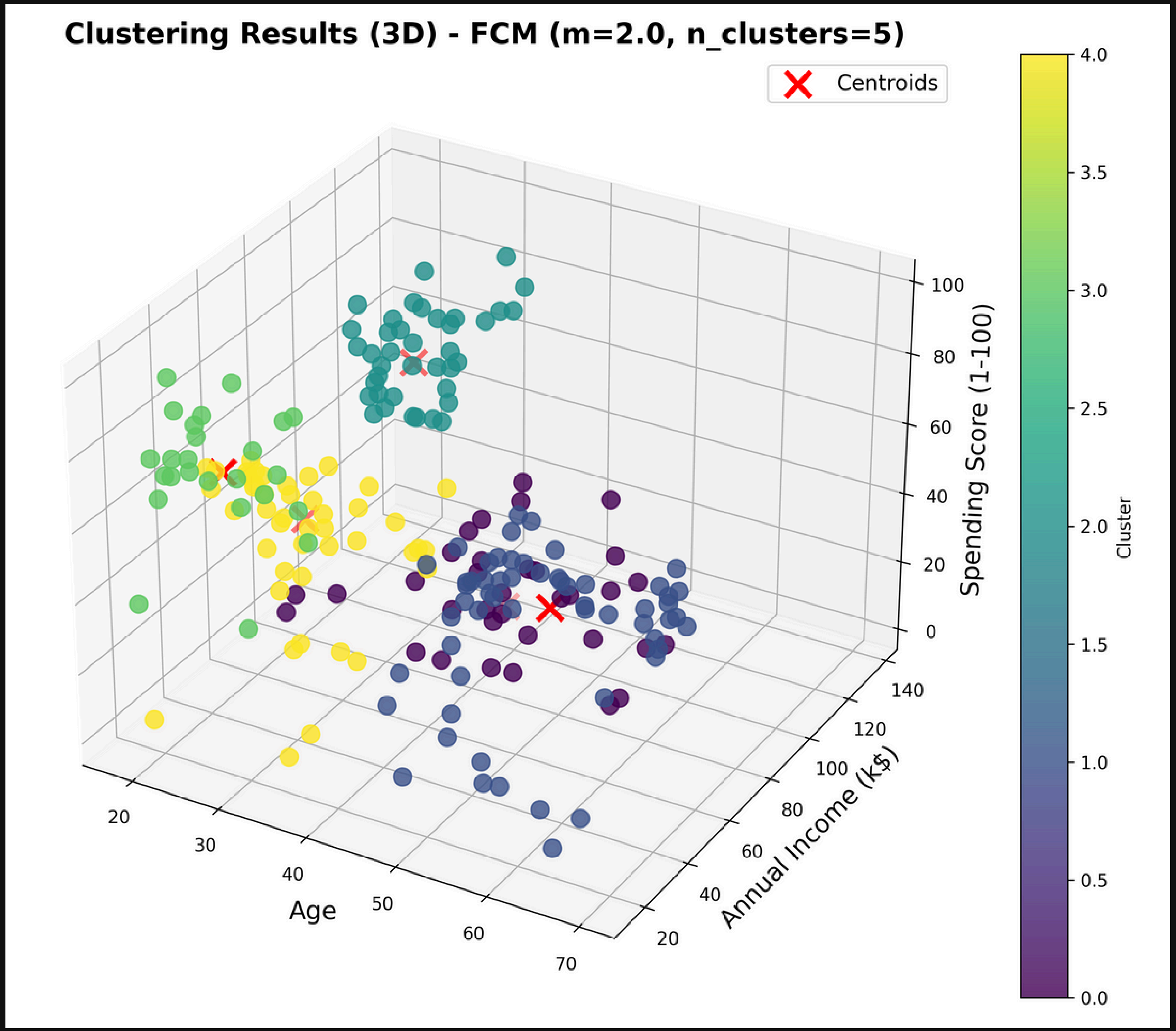


VISUALIZATION AND ANALYSIS

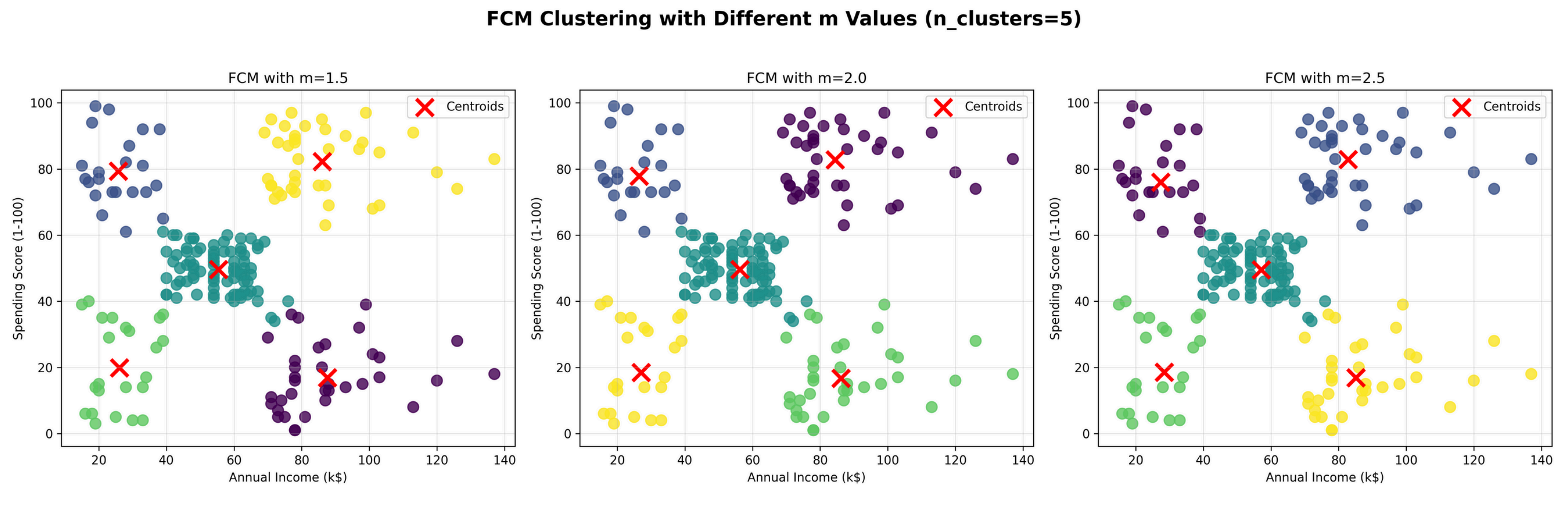


VISUALIZATION AND ANALYSIS

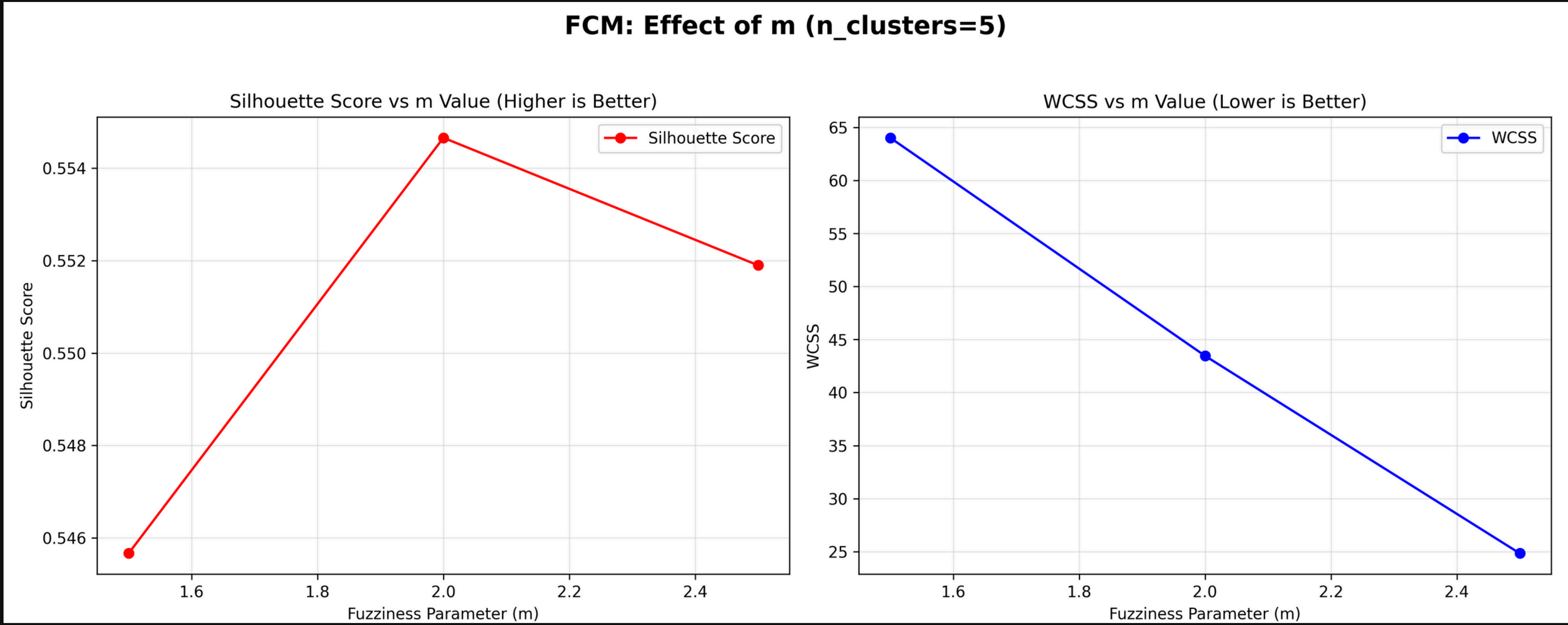




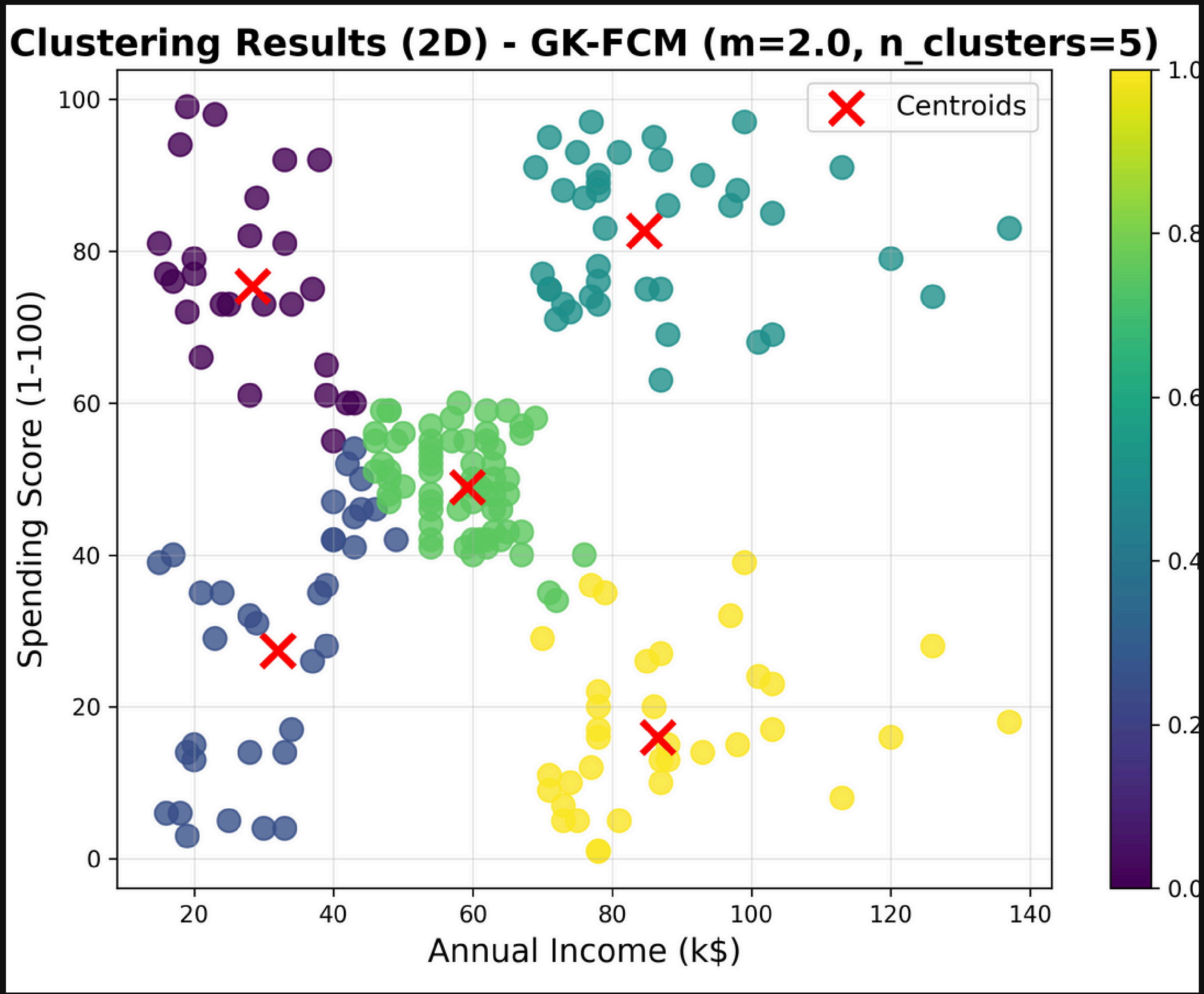
VISUALIZATION AND ANALYSIS

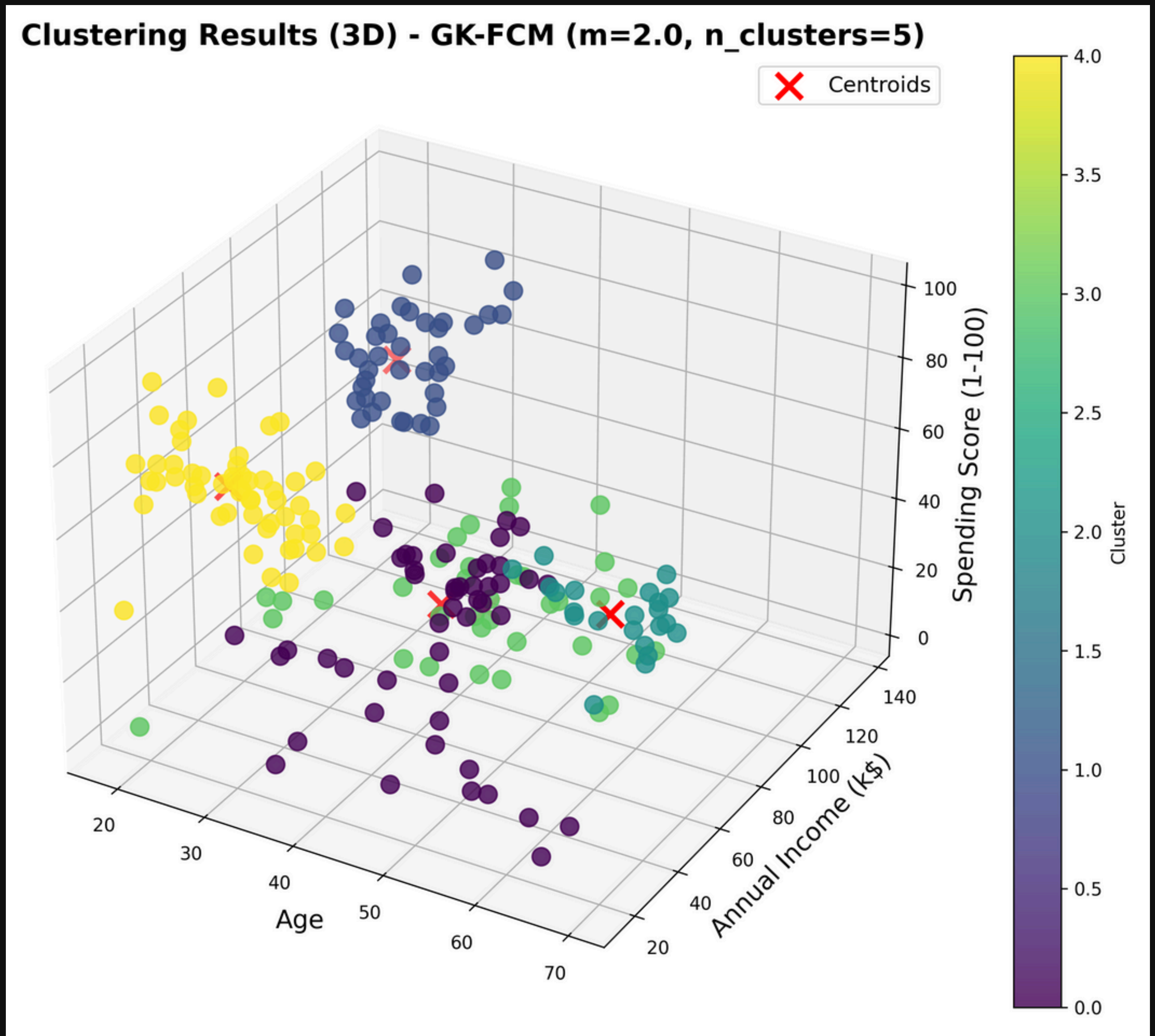


VISUALIZATION AND ANALYSIS

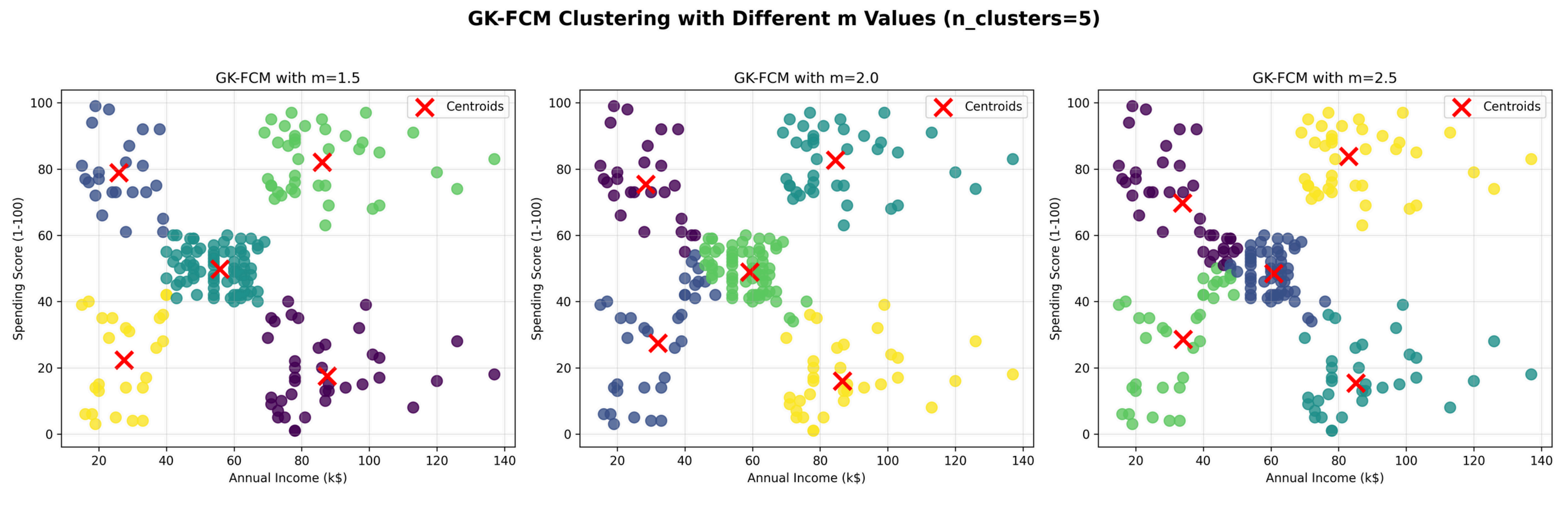


VISUALIZATION AND ANALYSIS

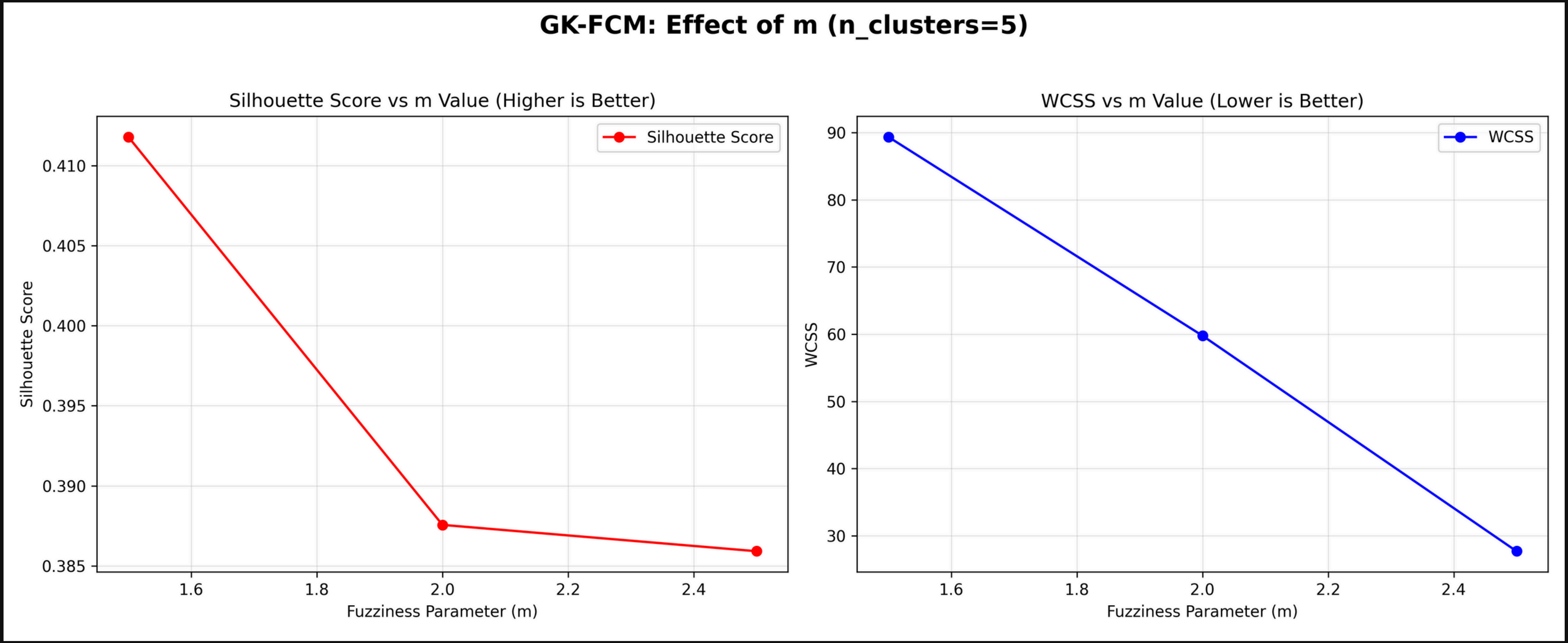




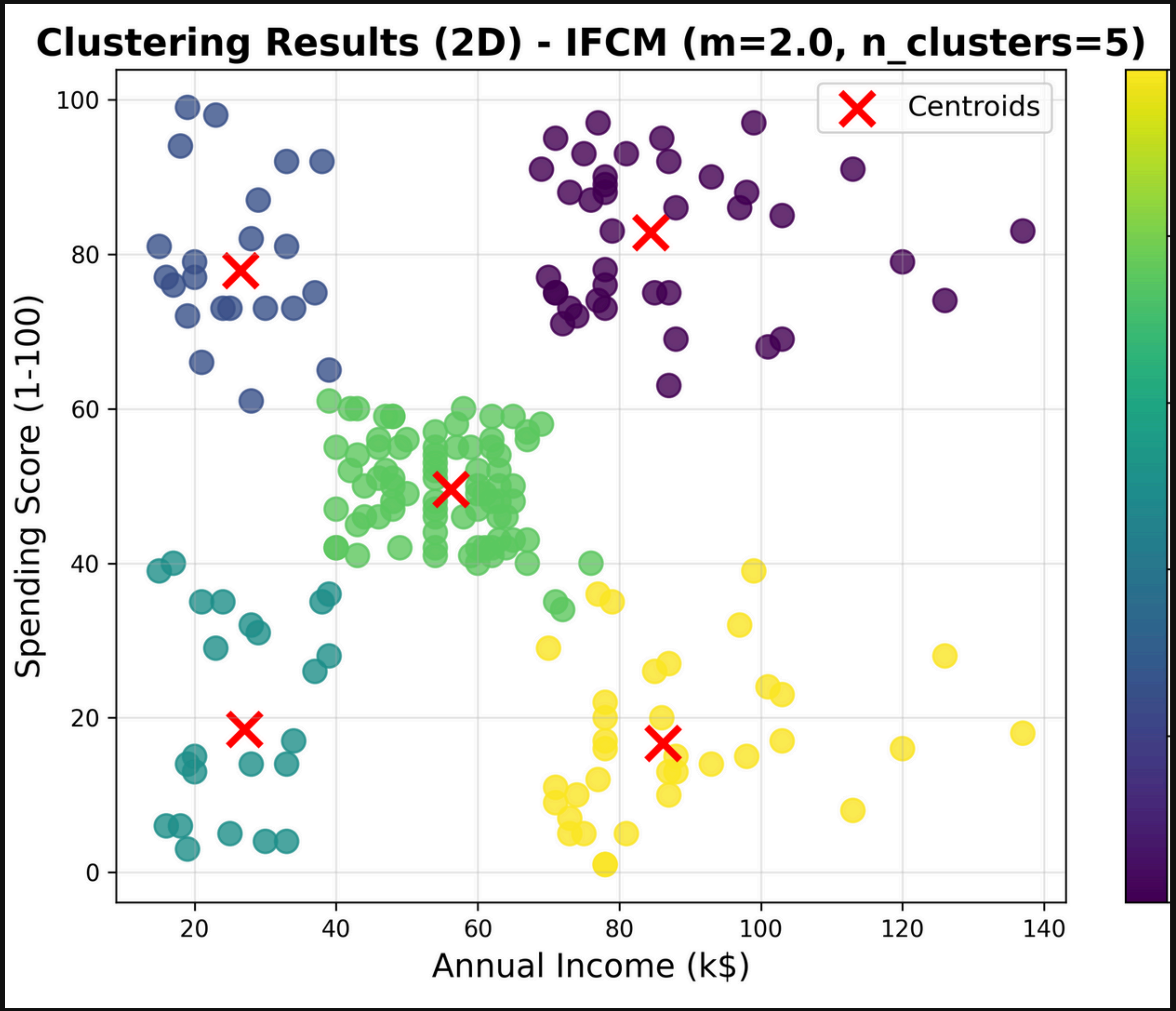
VISUALIZATION AND ANALYSIS

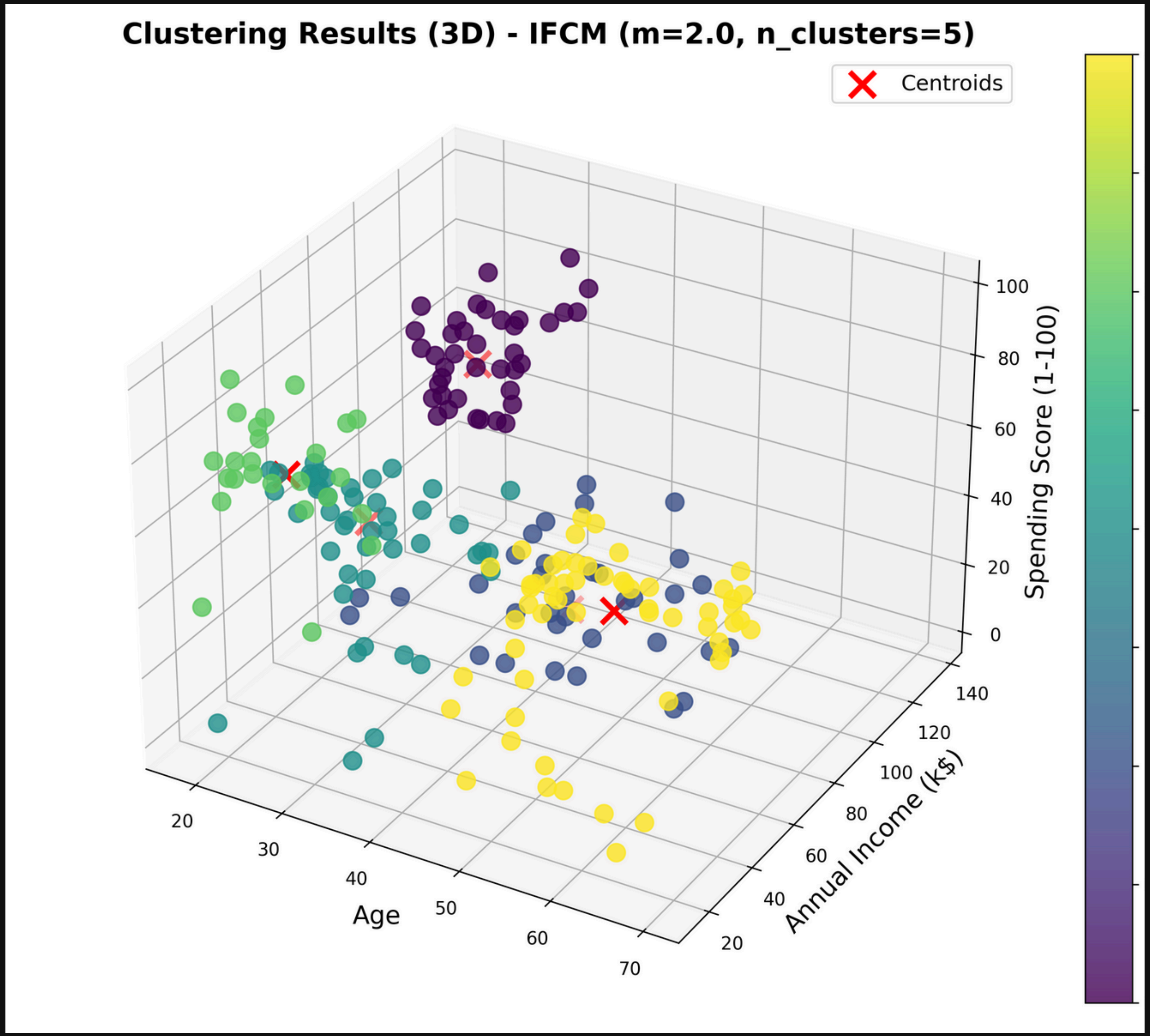


VISUALIZATION AND ANALYSIS

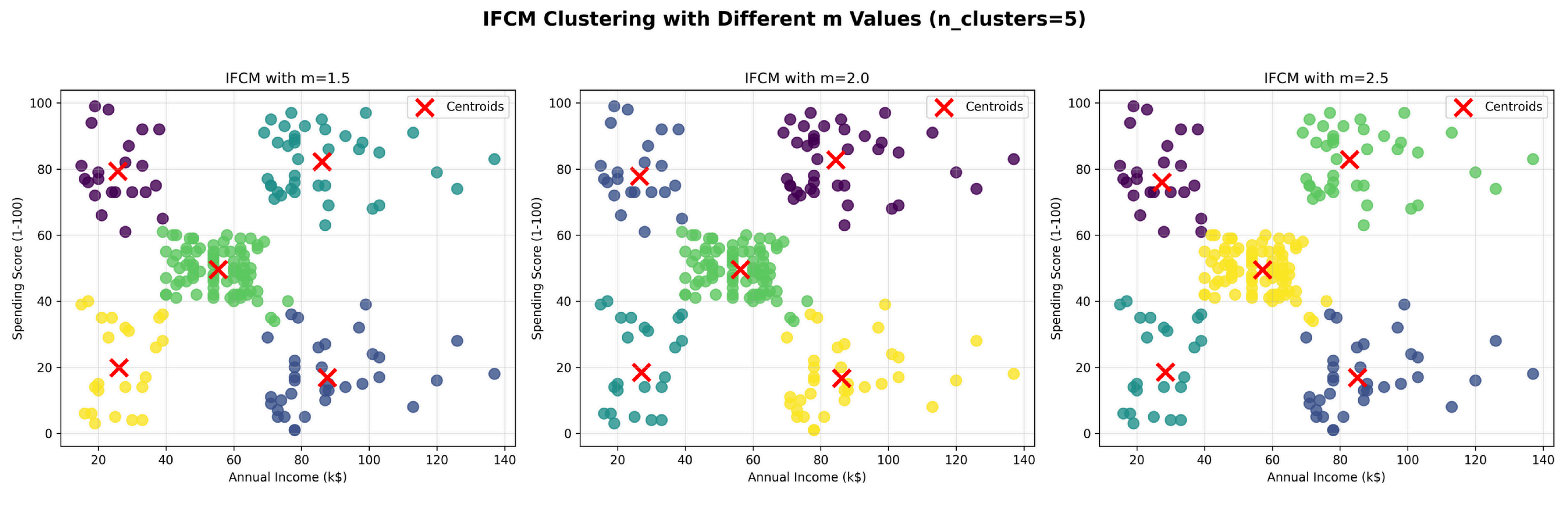


VISUALIZATION AND ANALYSIS

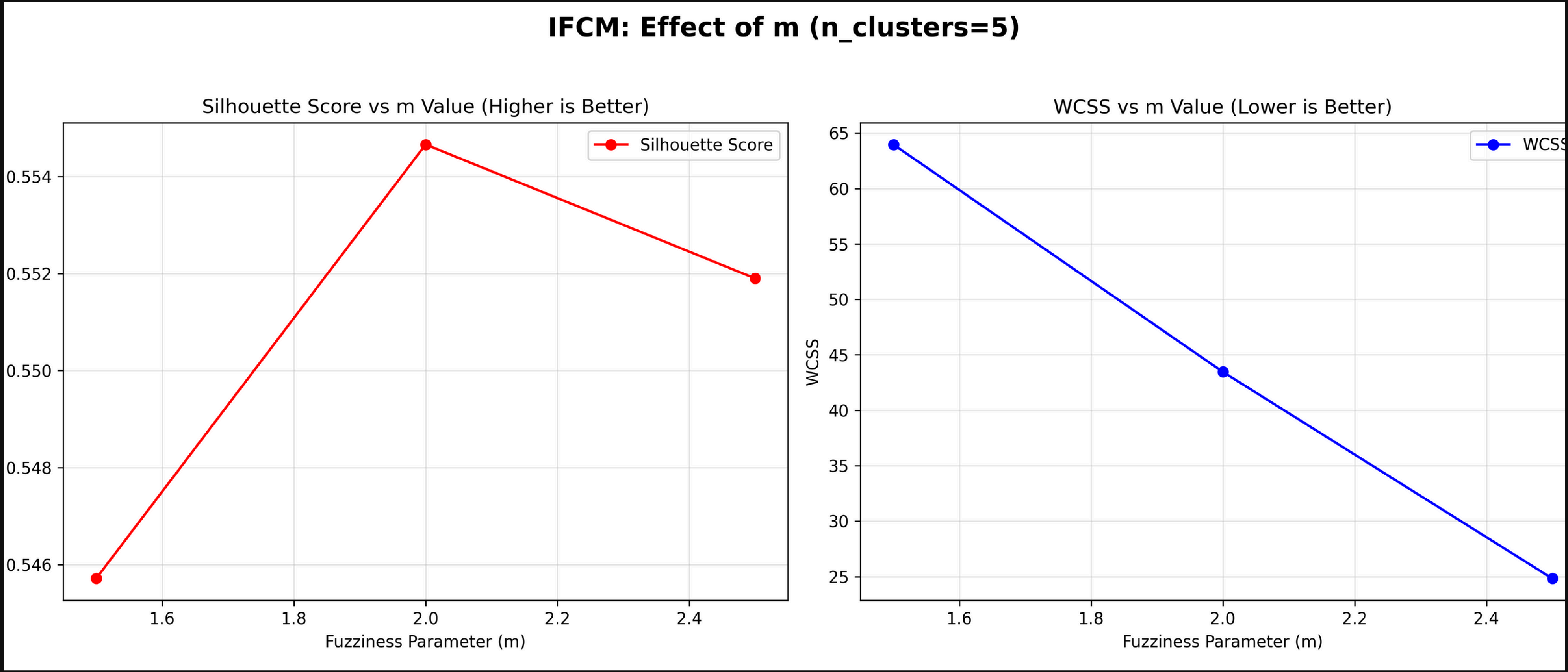




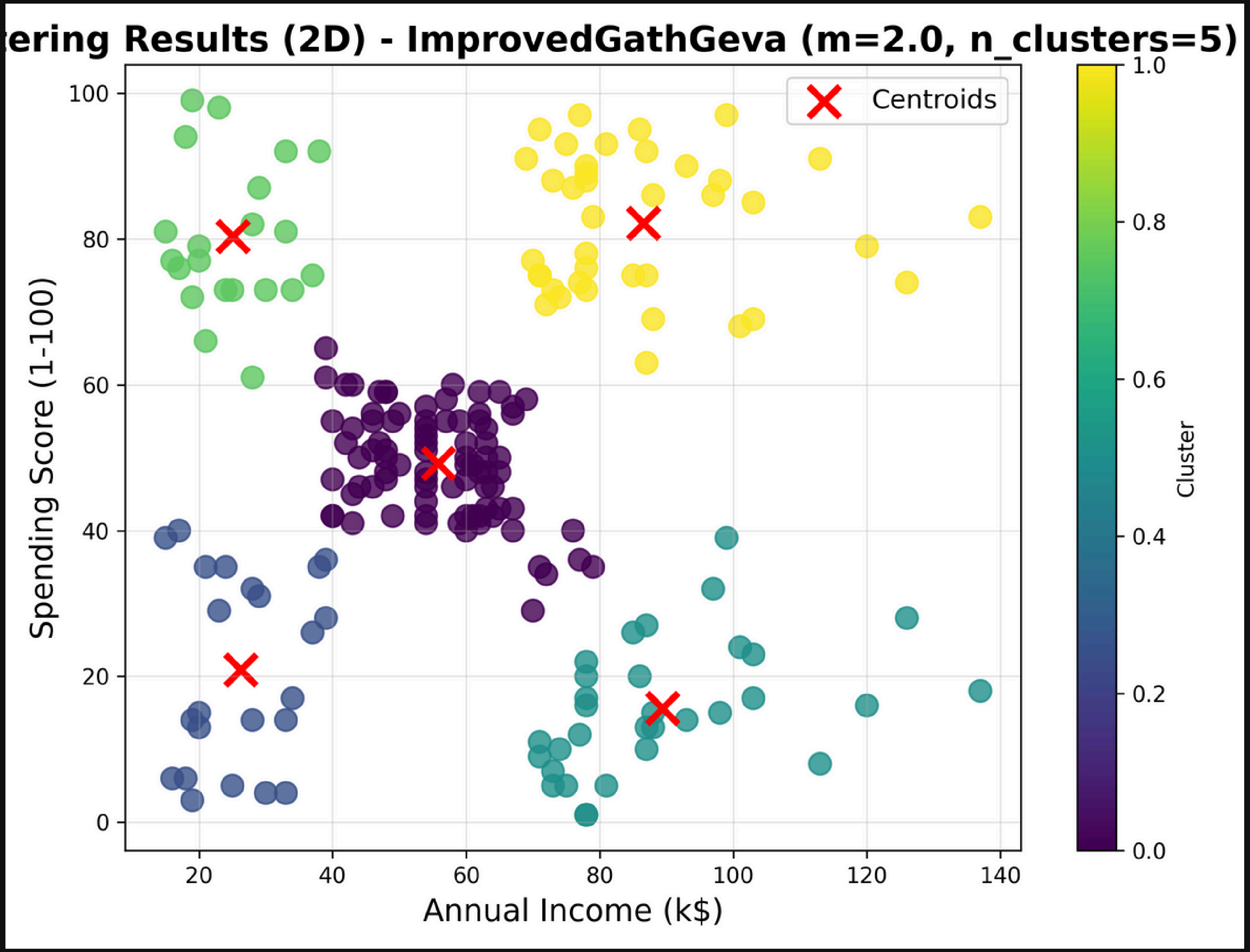
VISUALIZATION AND ANALYSIS

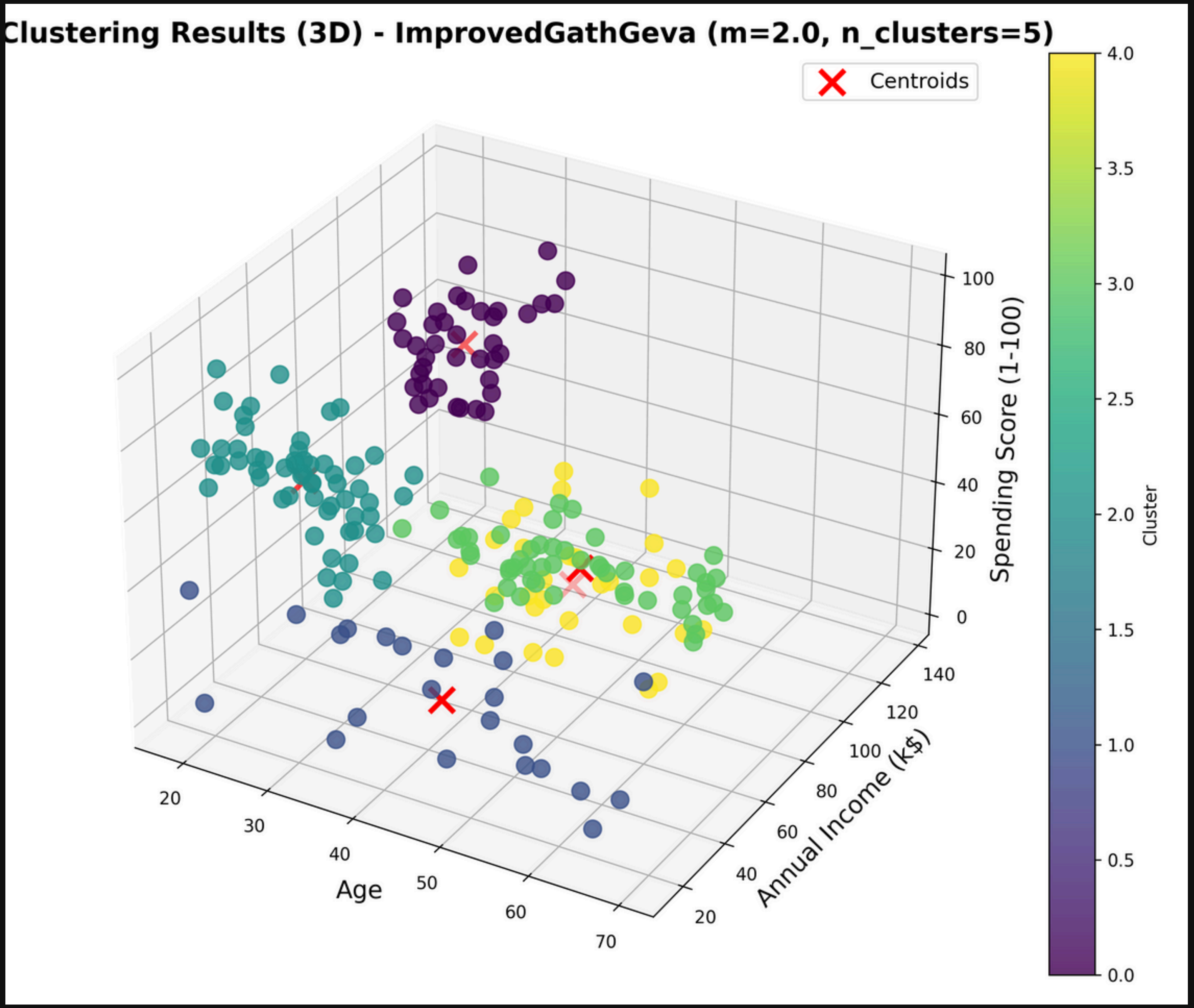


VISUALIZATION AND ANALYSIS

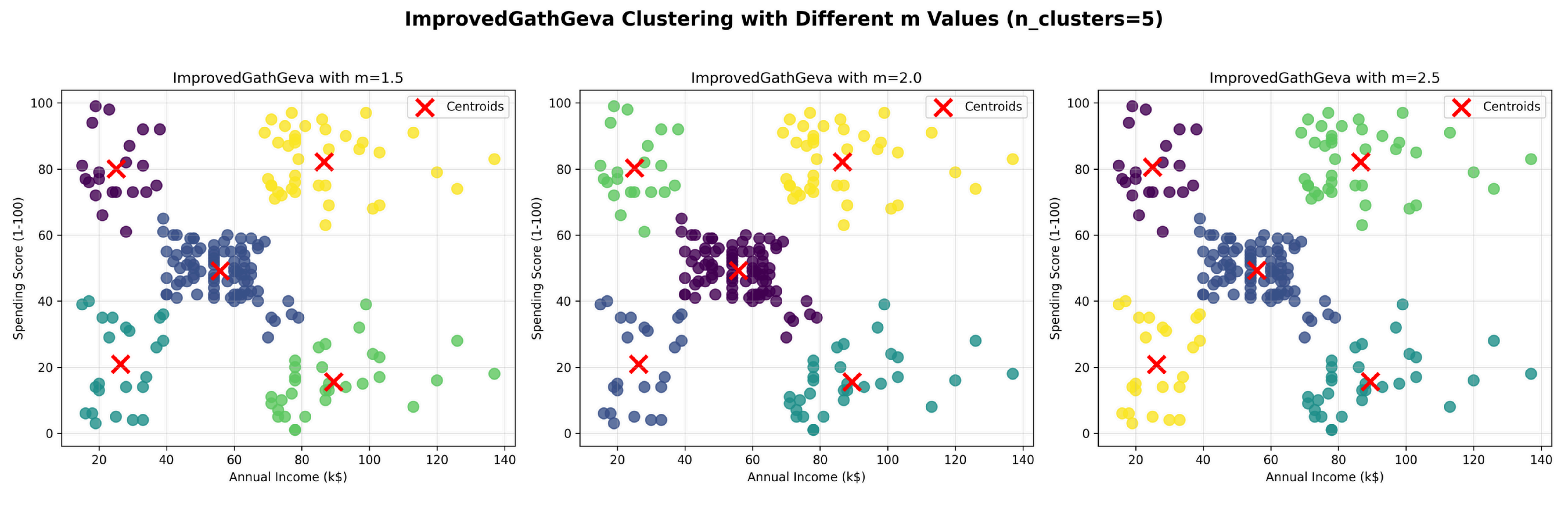


VISUALIZATION AND ANALYSIS

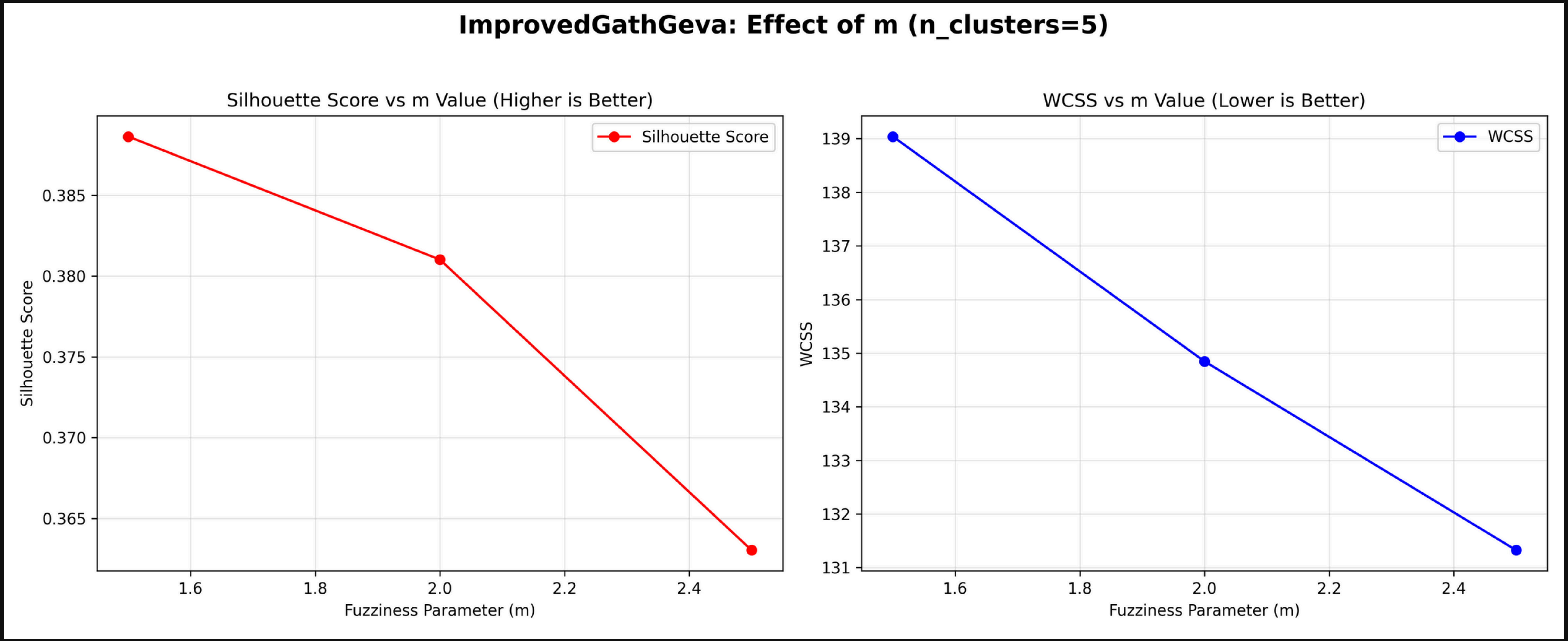




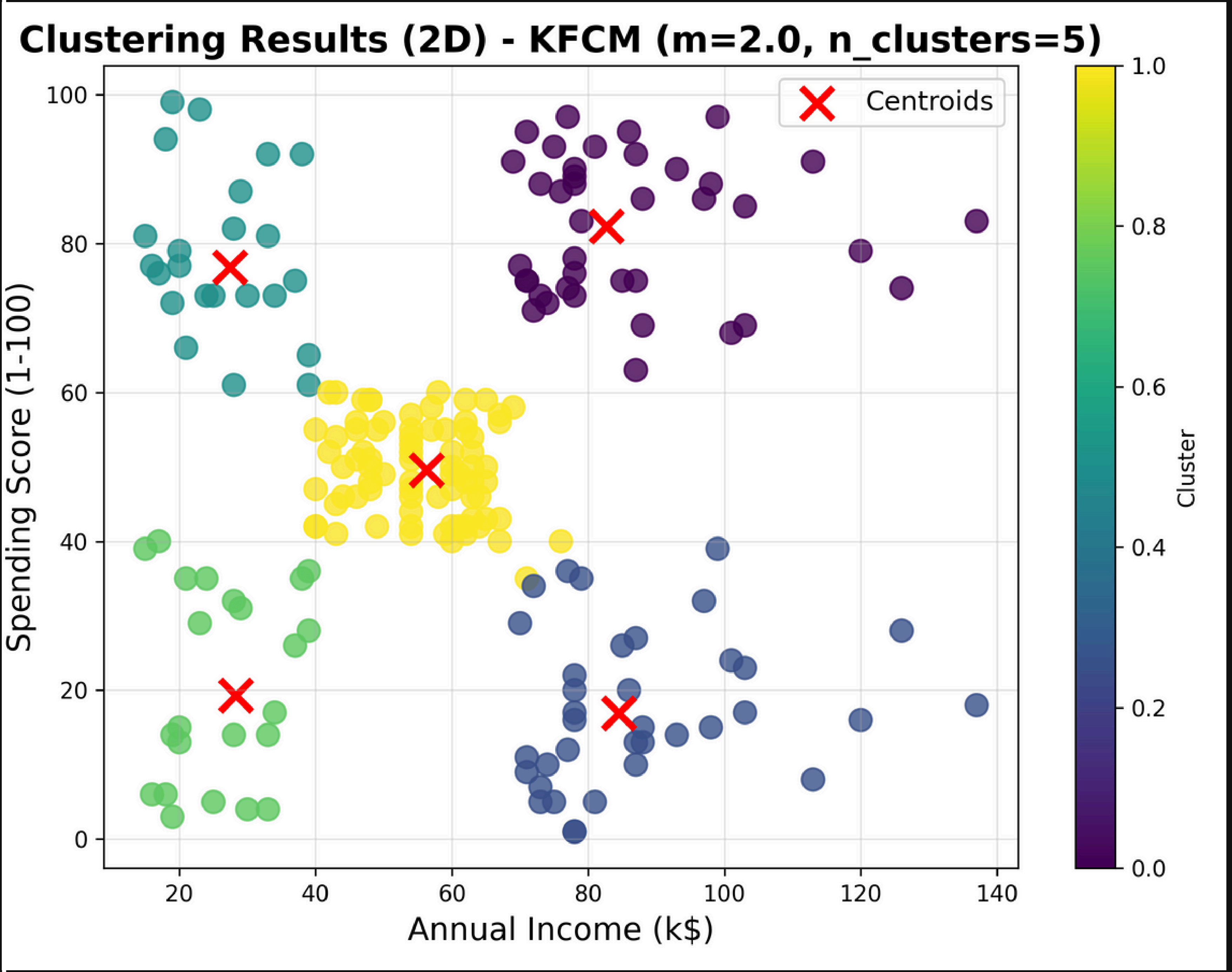
VISUALIZATION AND ANALYSIS

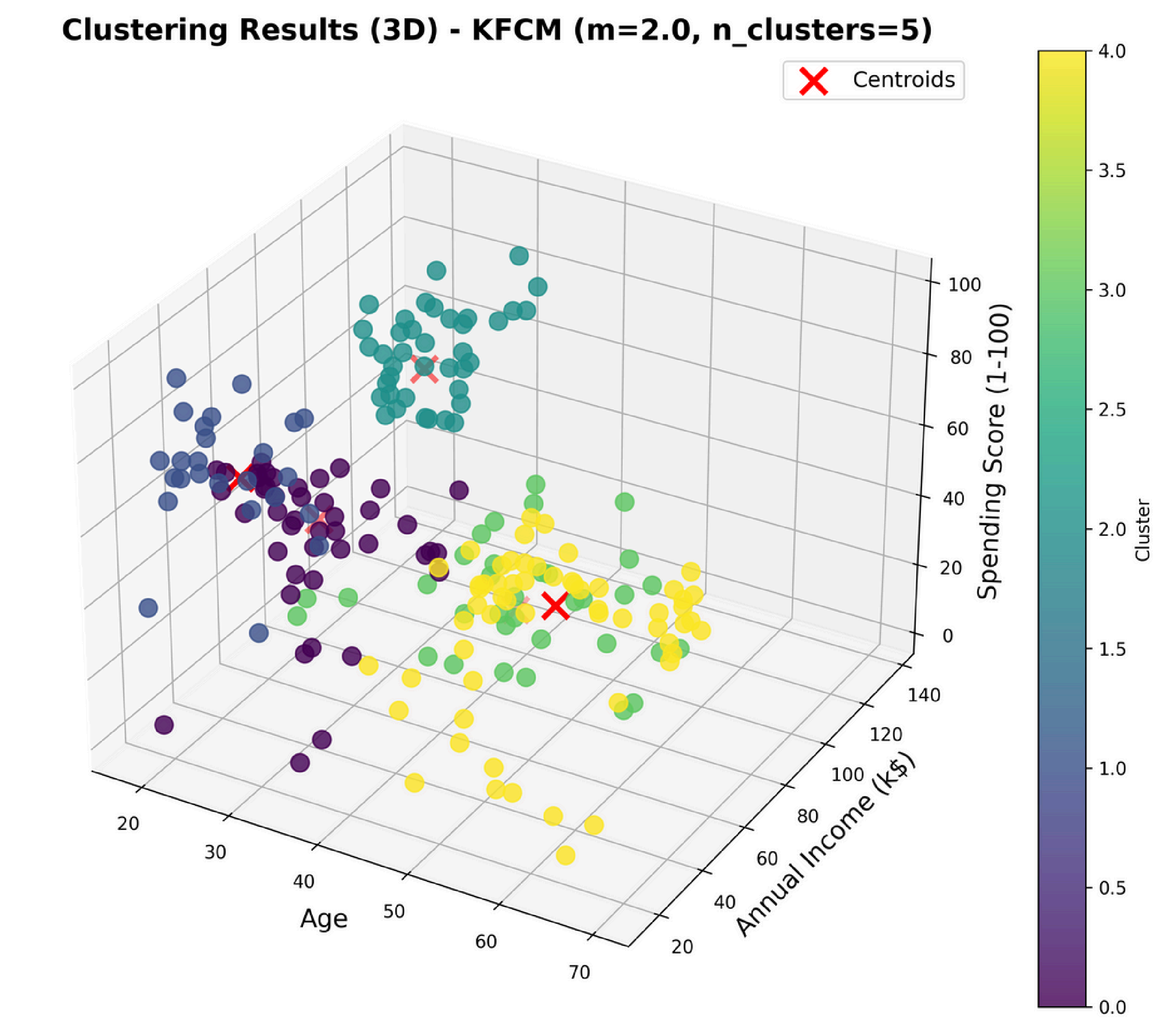


VISUALIZATION AND ANALYSIS

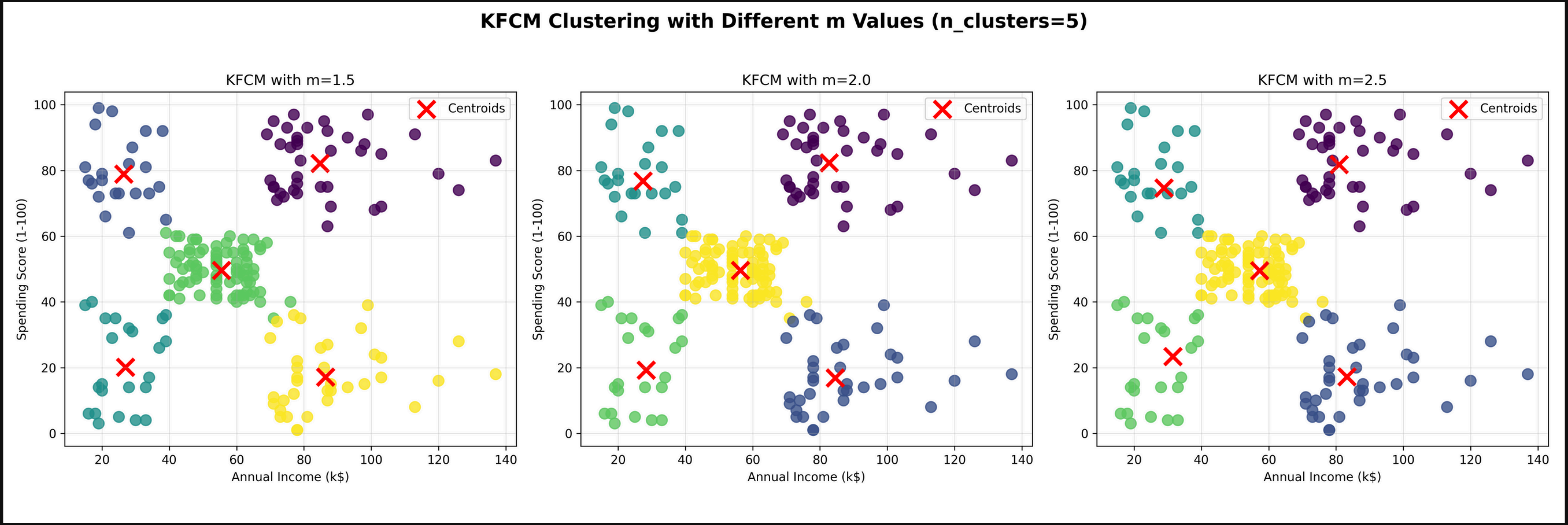


VISUALIZATION AND ANALYSIS

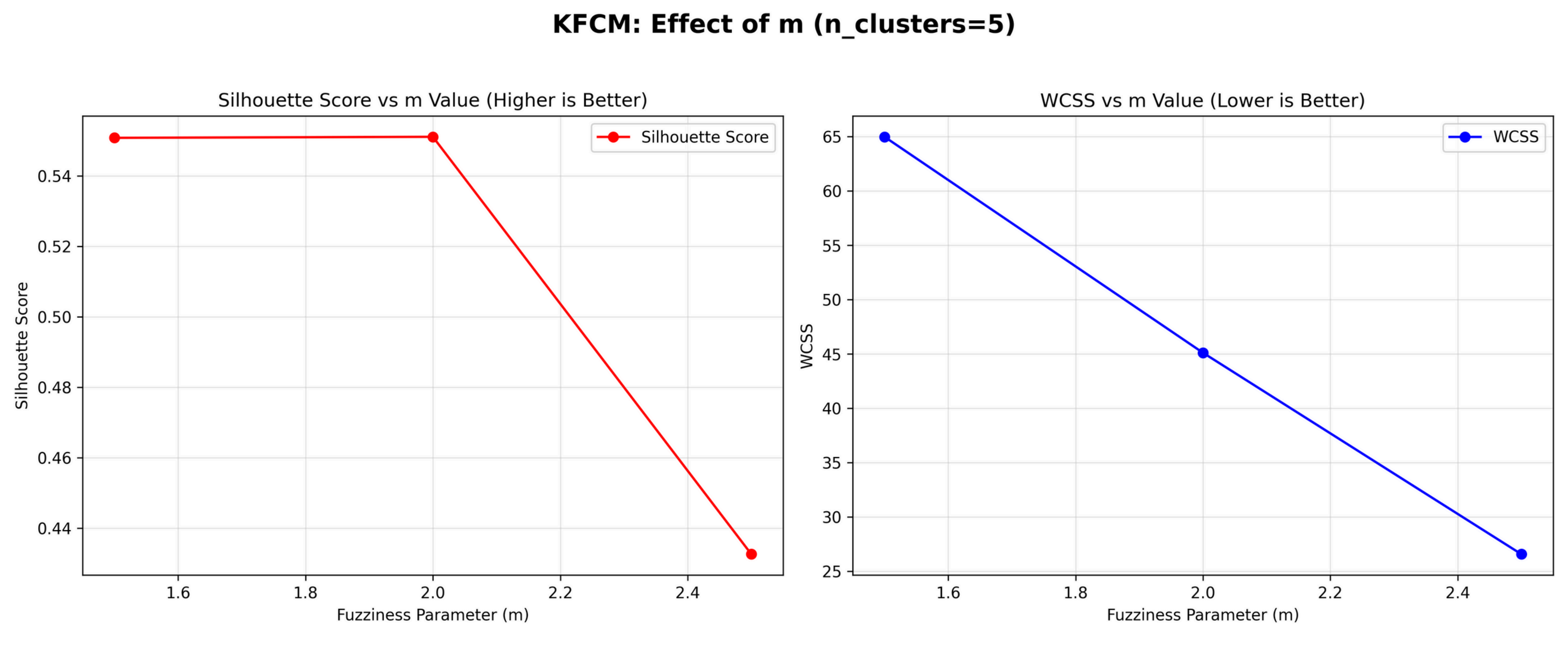




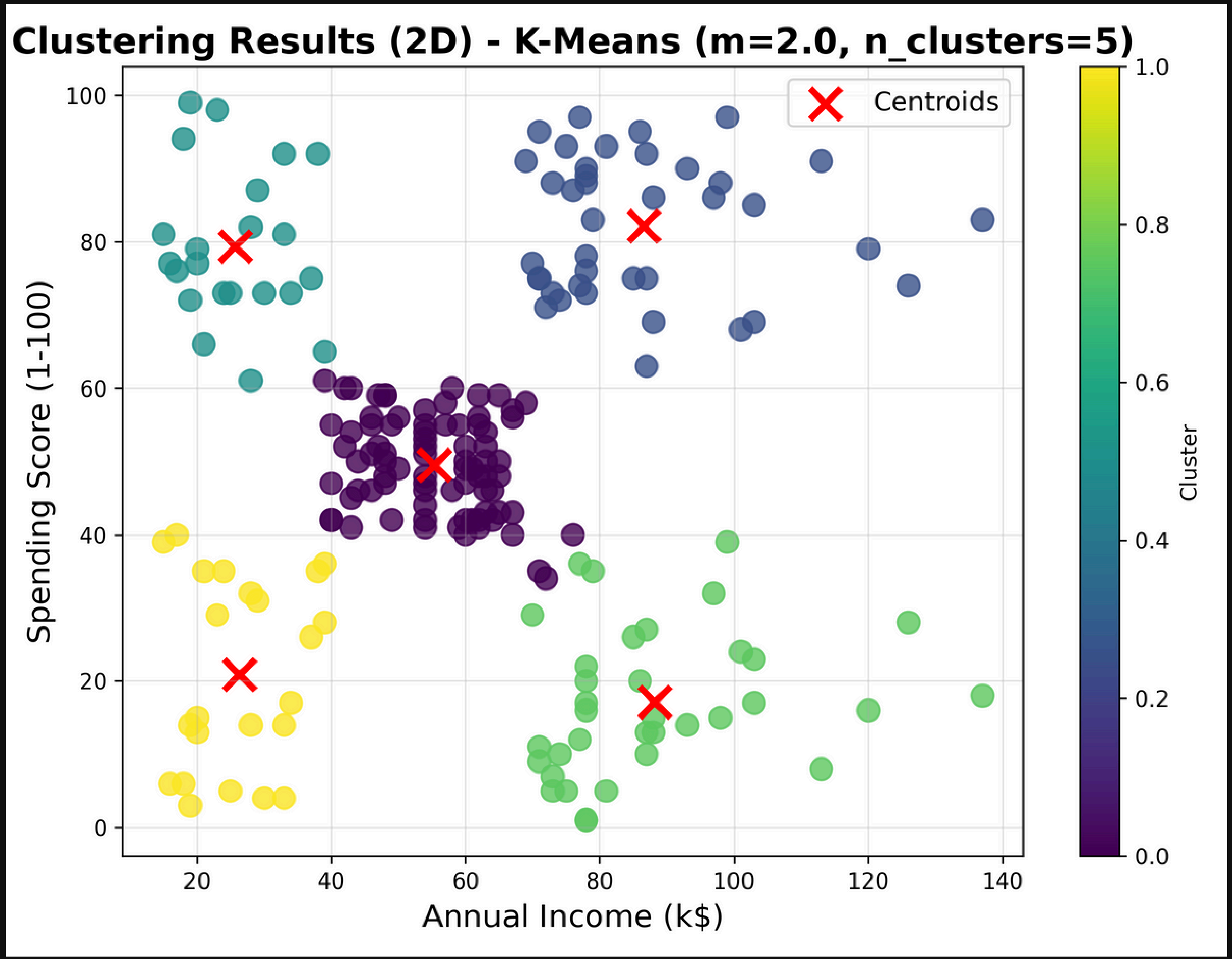
VISUALIZATION AND ANALYSIS

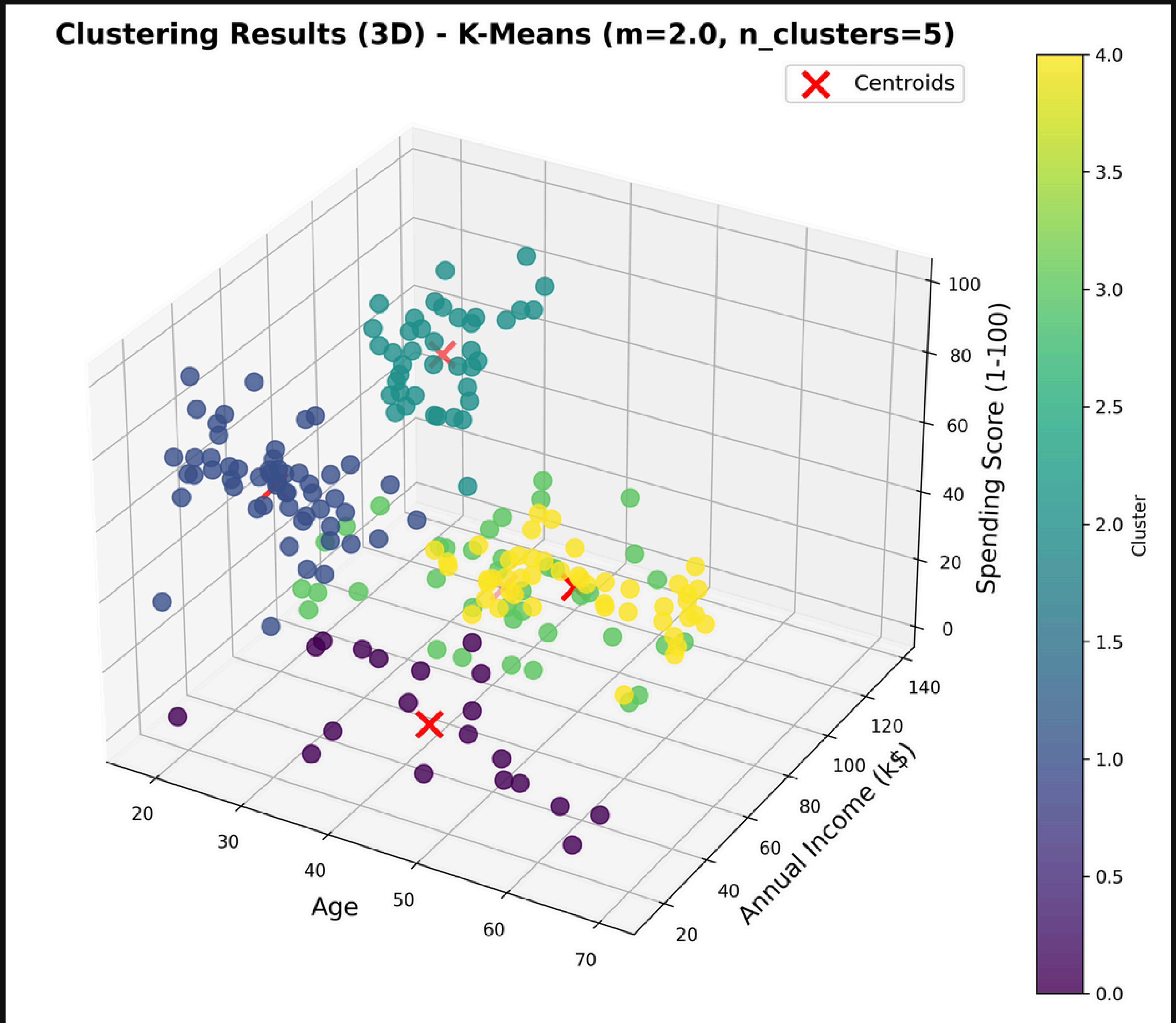


VISUALIZATION AND ANALYSIS

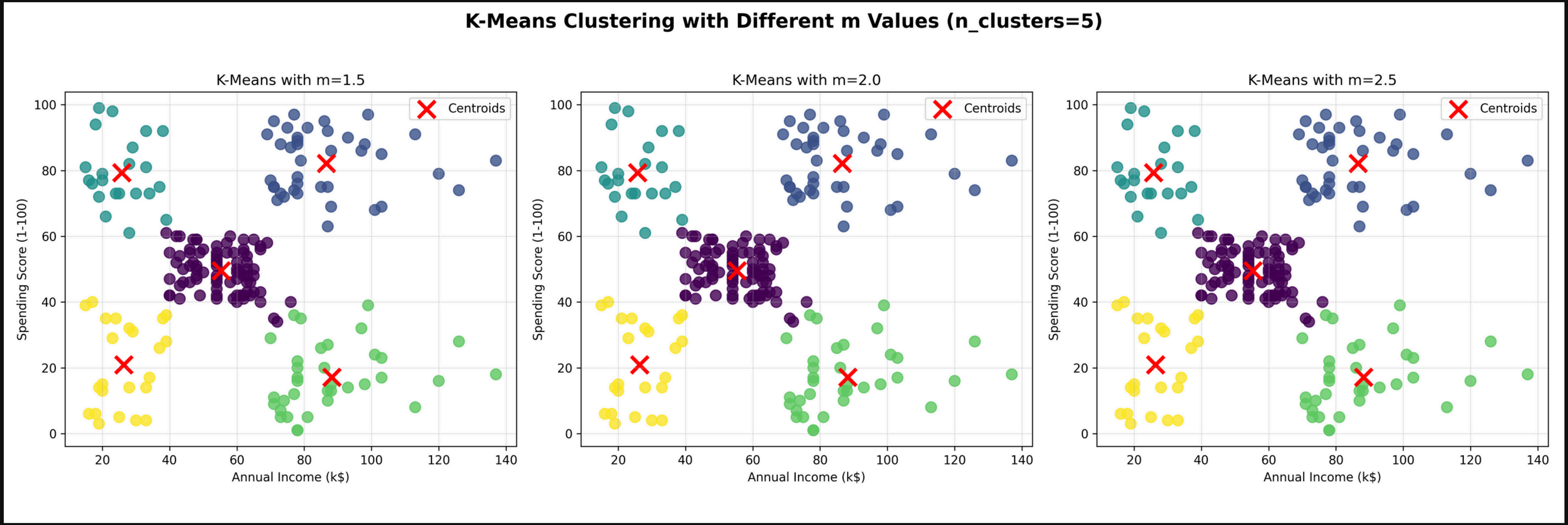


VISUALIZATION AND ANALYSIS

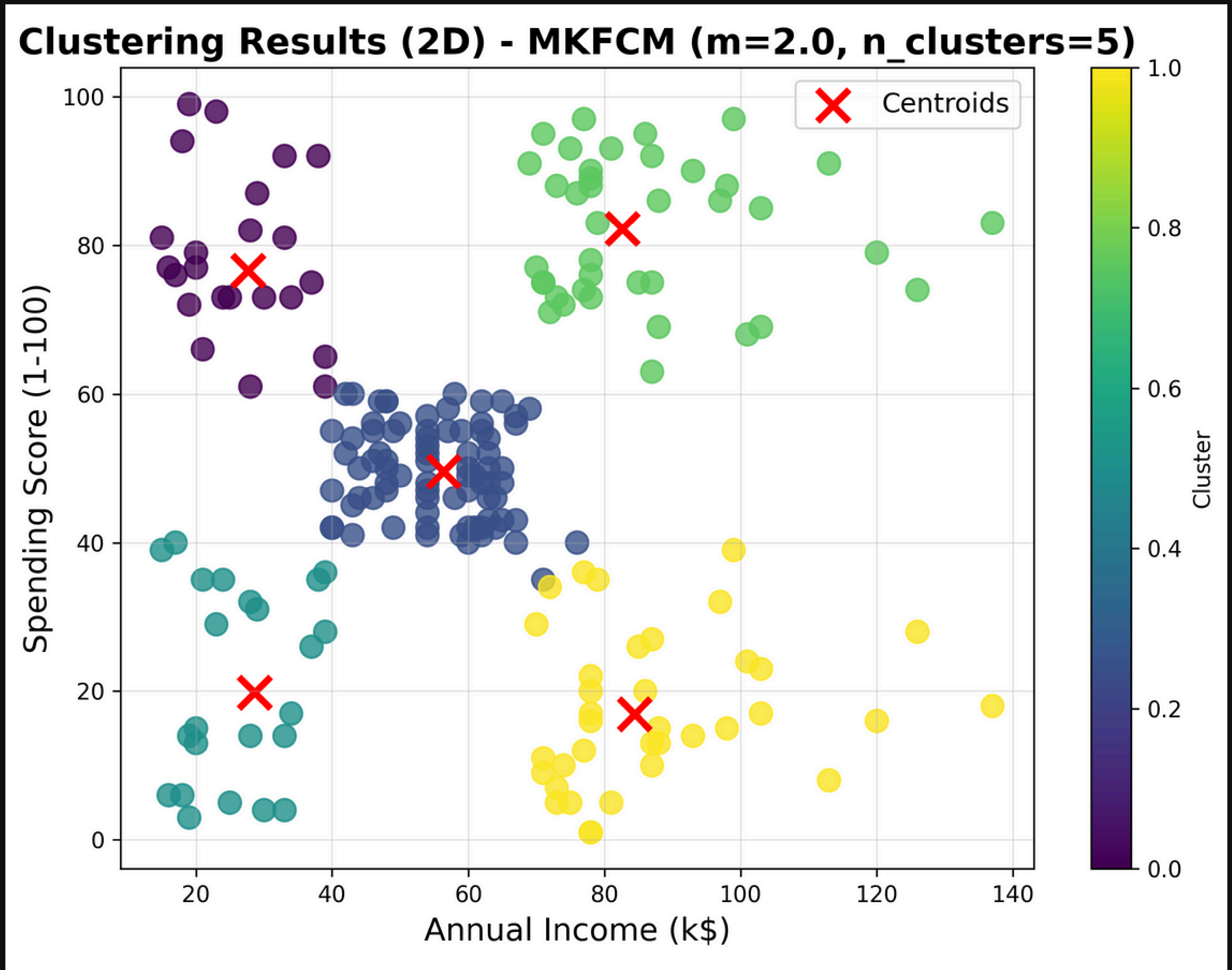


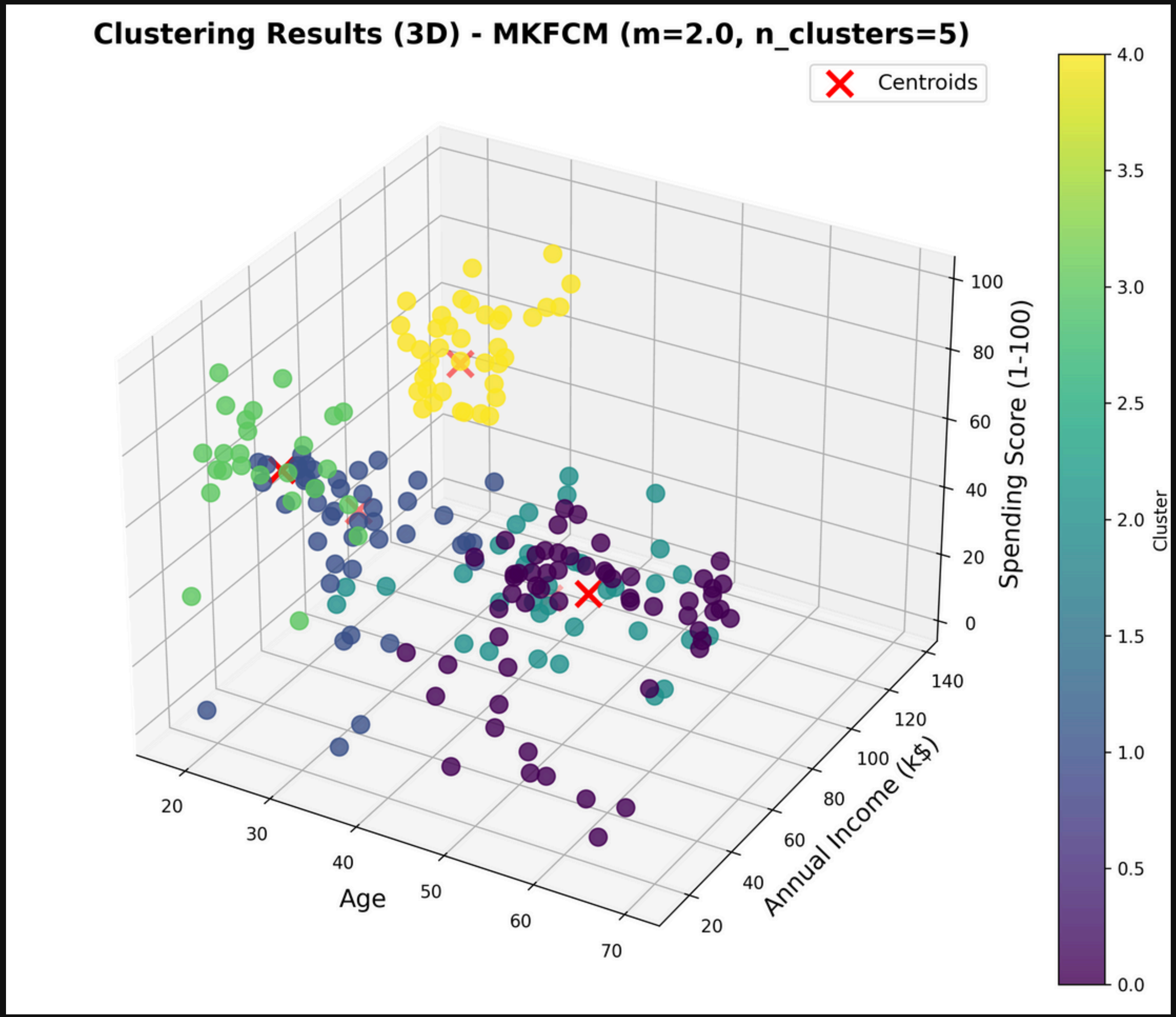


VISUALIZATION AND ANALYSIS

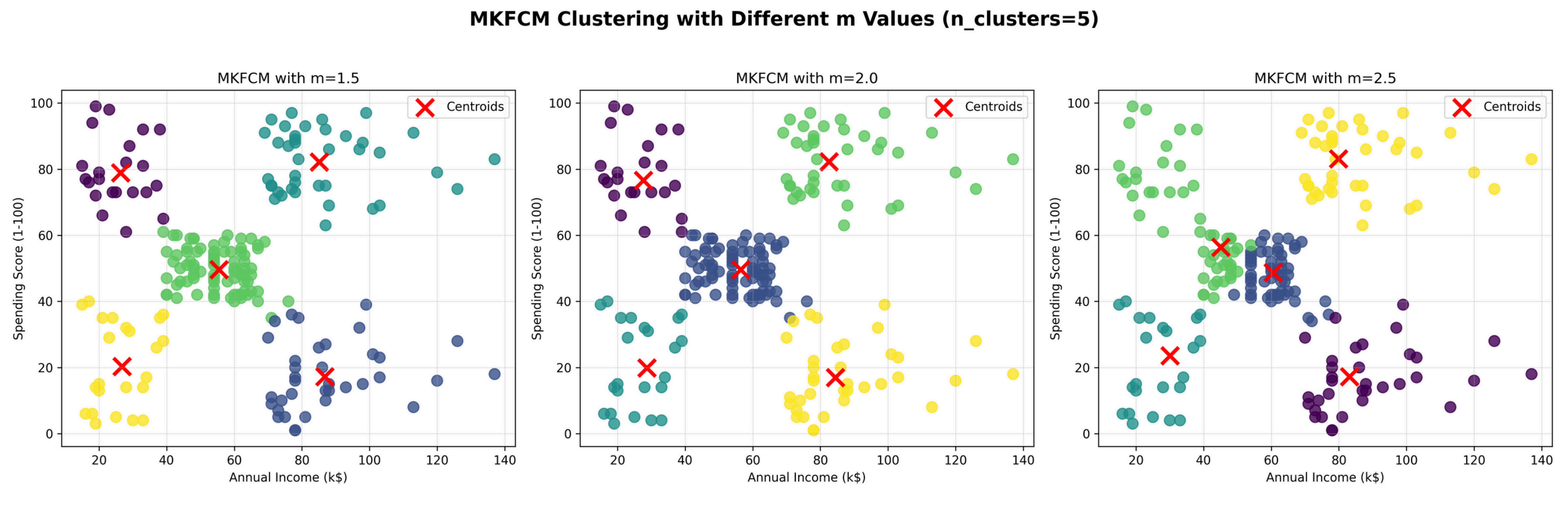


VISUALIZATION AND ANALYSIS

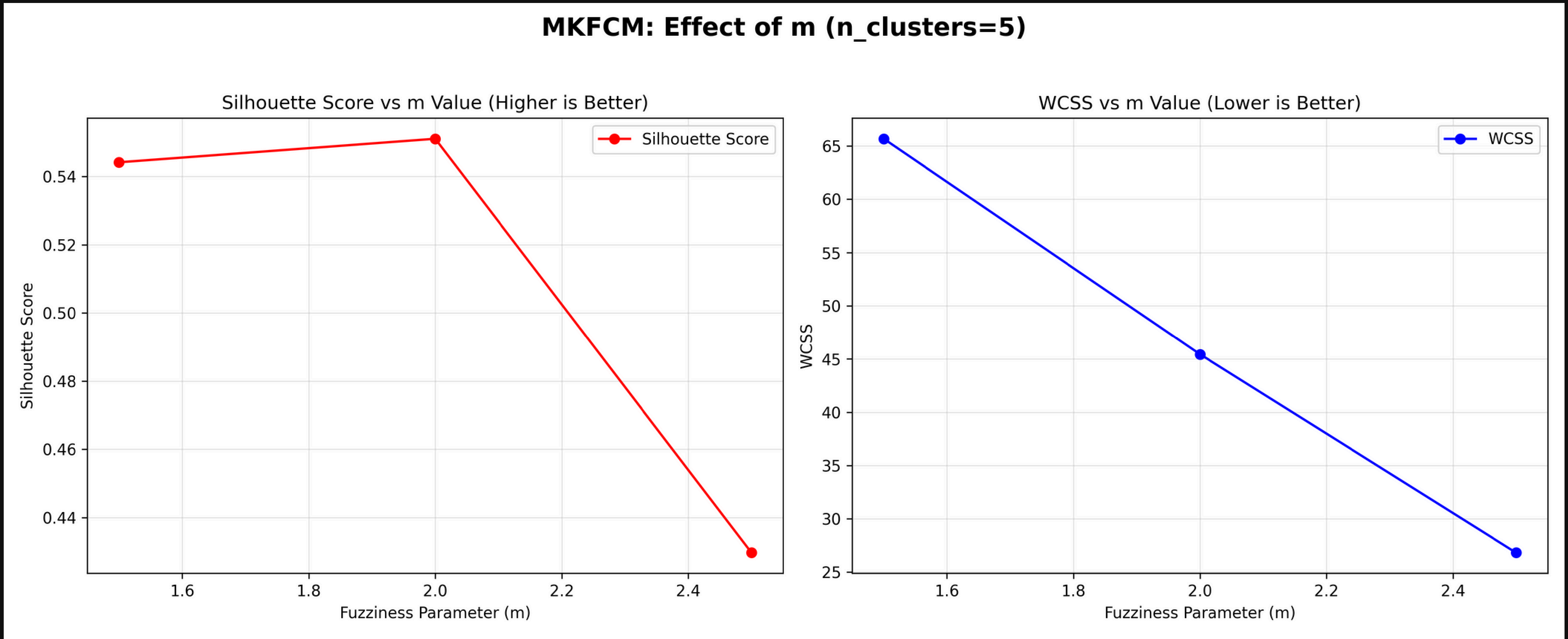




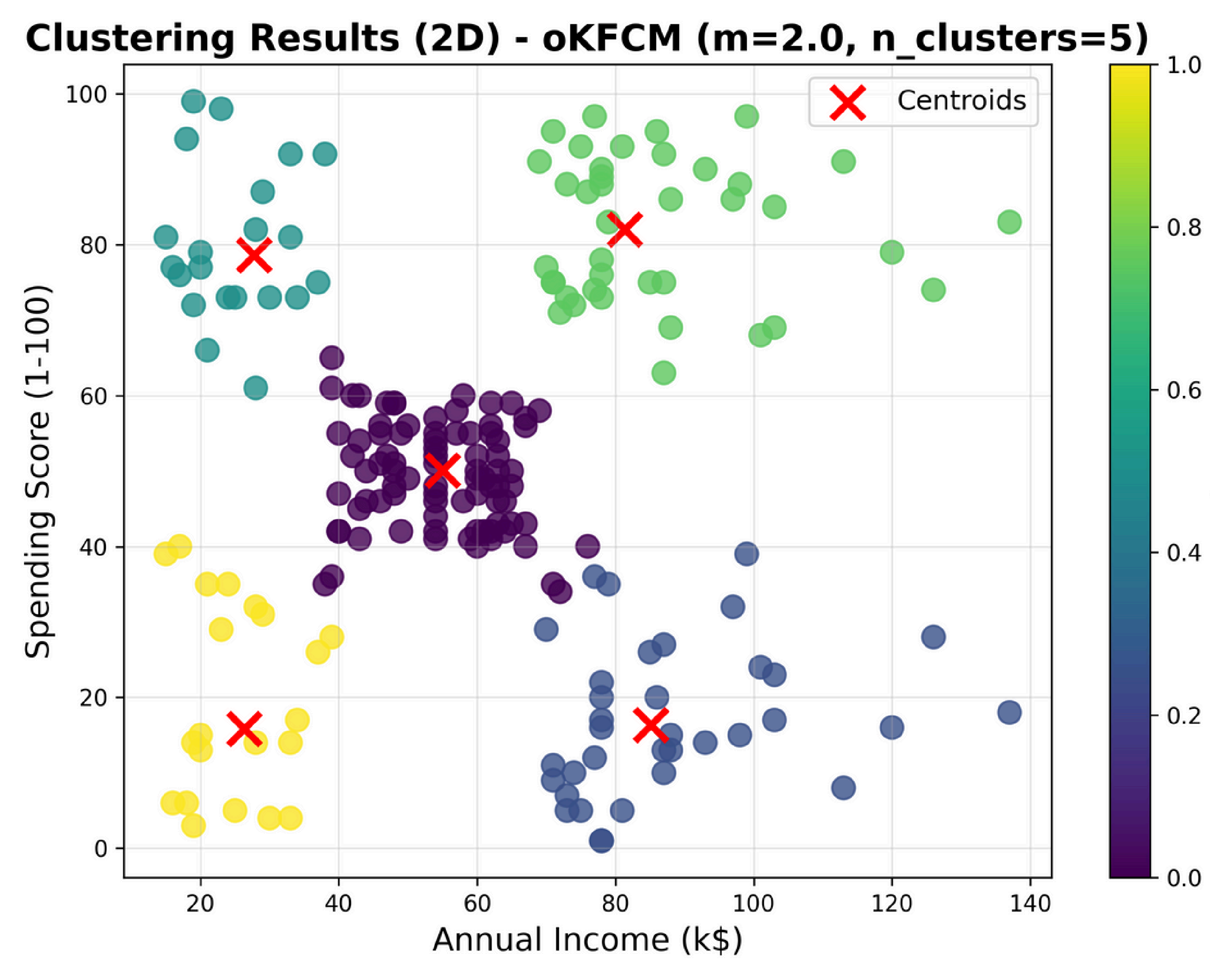
VISUALIZATION AND ANALYSIS

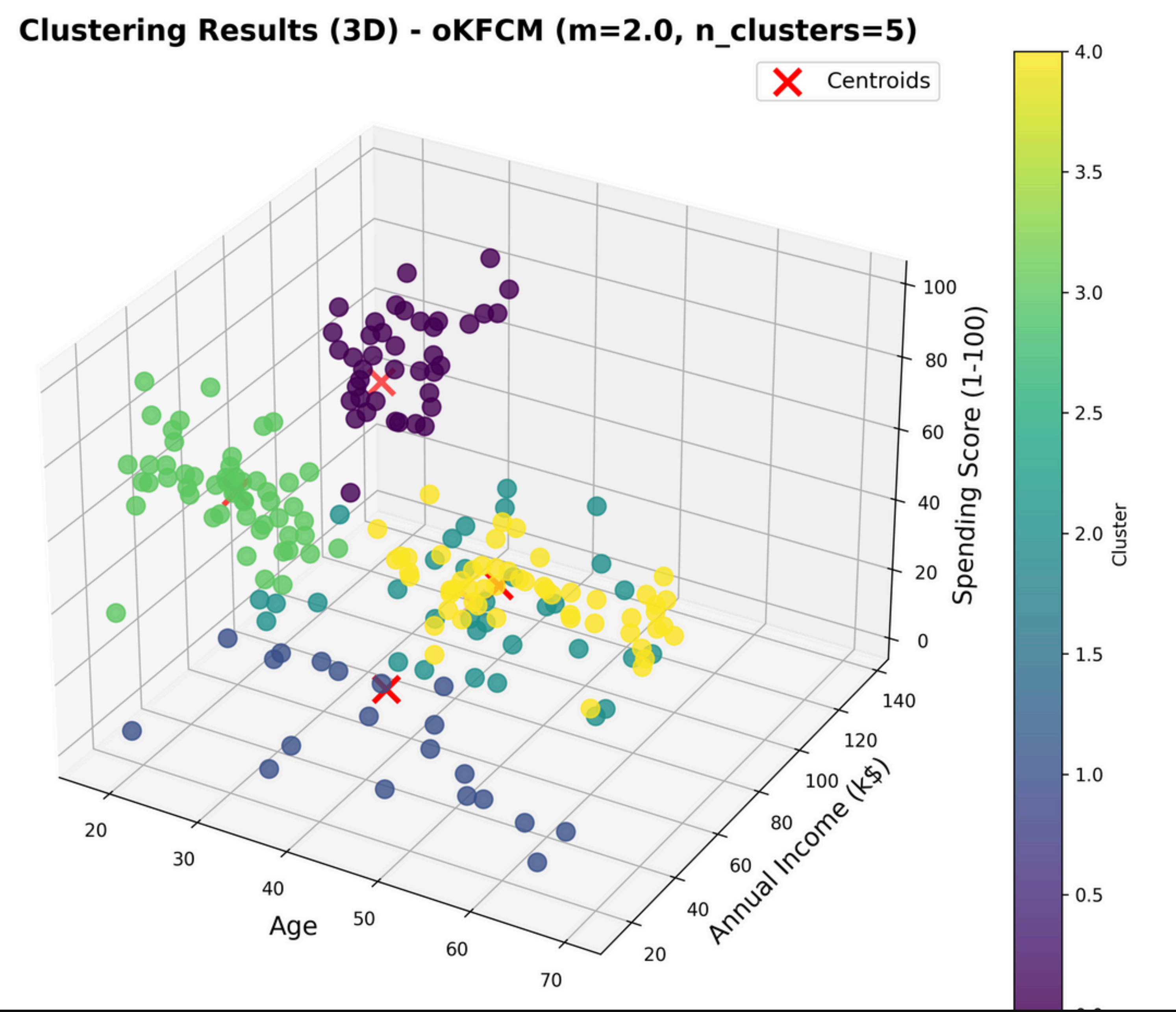


VISUALIZATION AND ANALYSIS

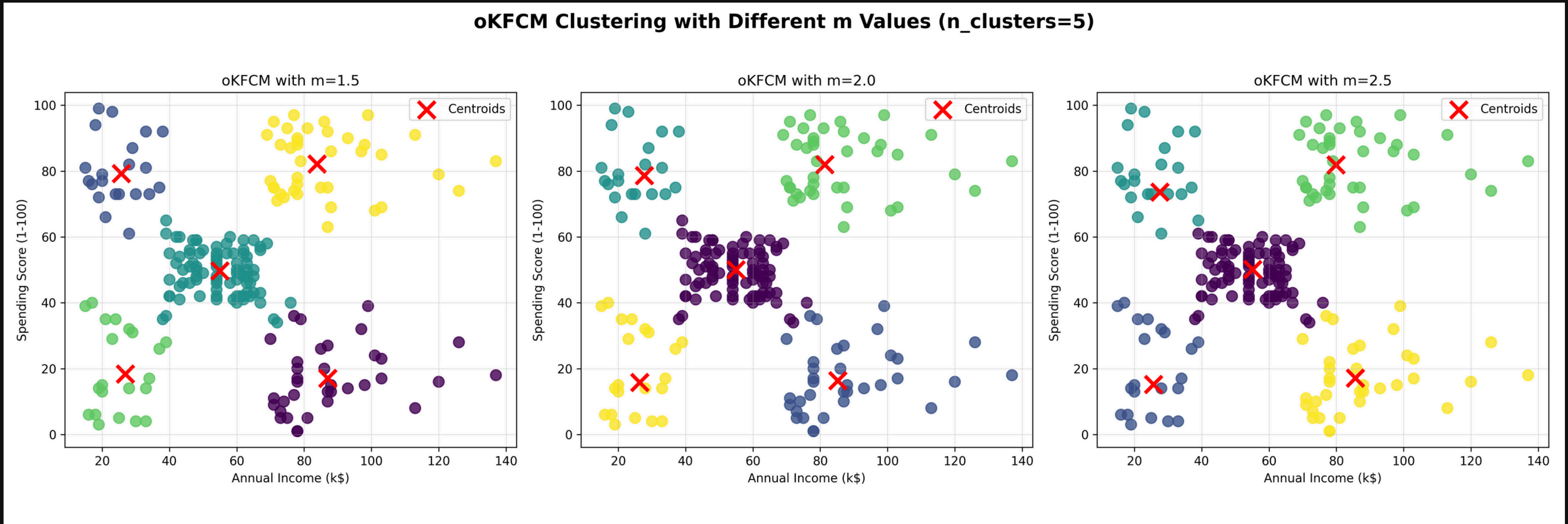


VISUALIZATION AND ANALYSIS

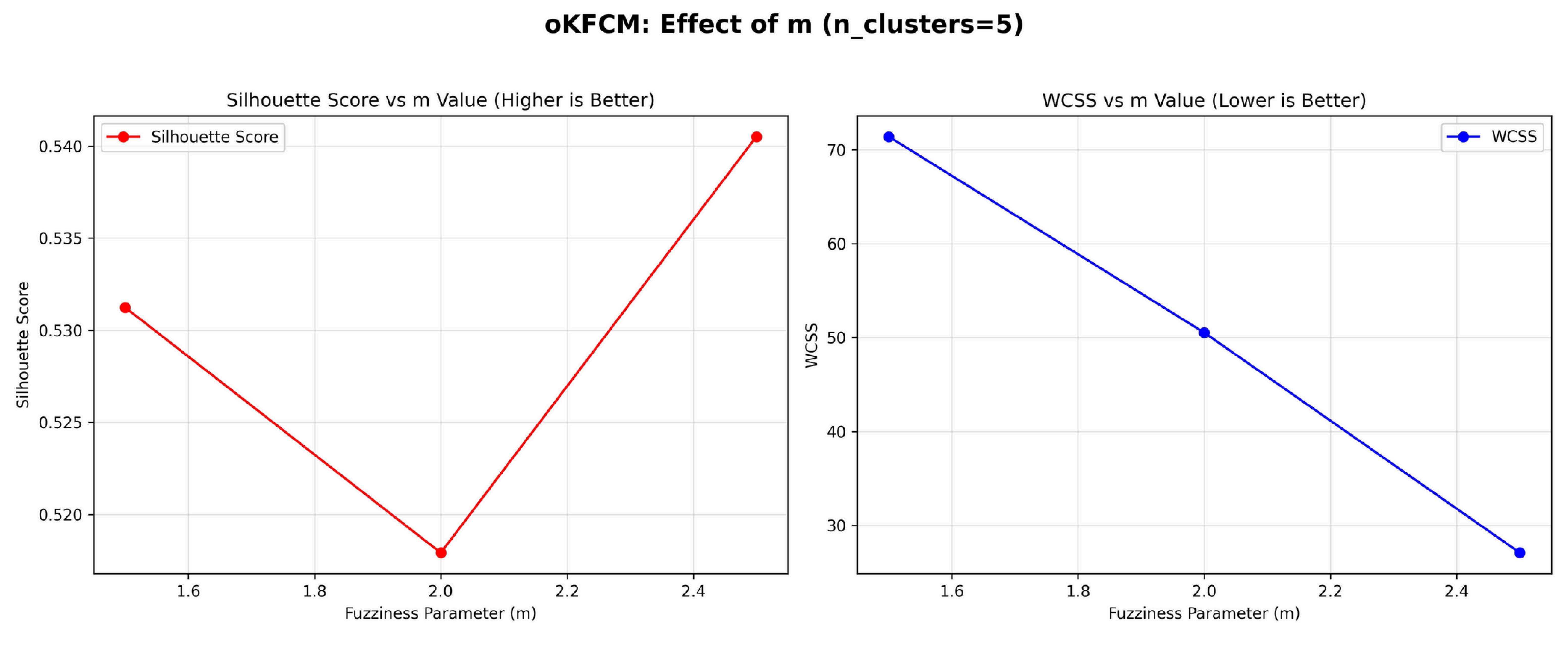




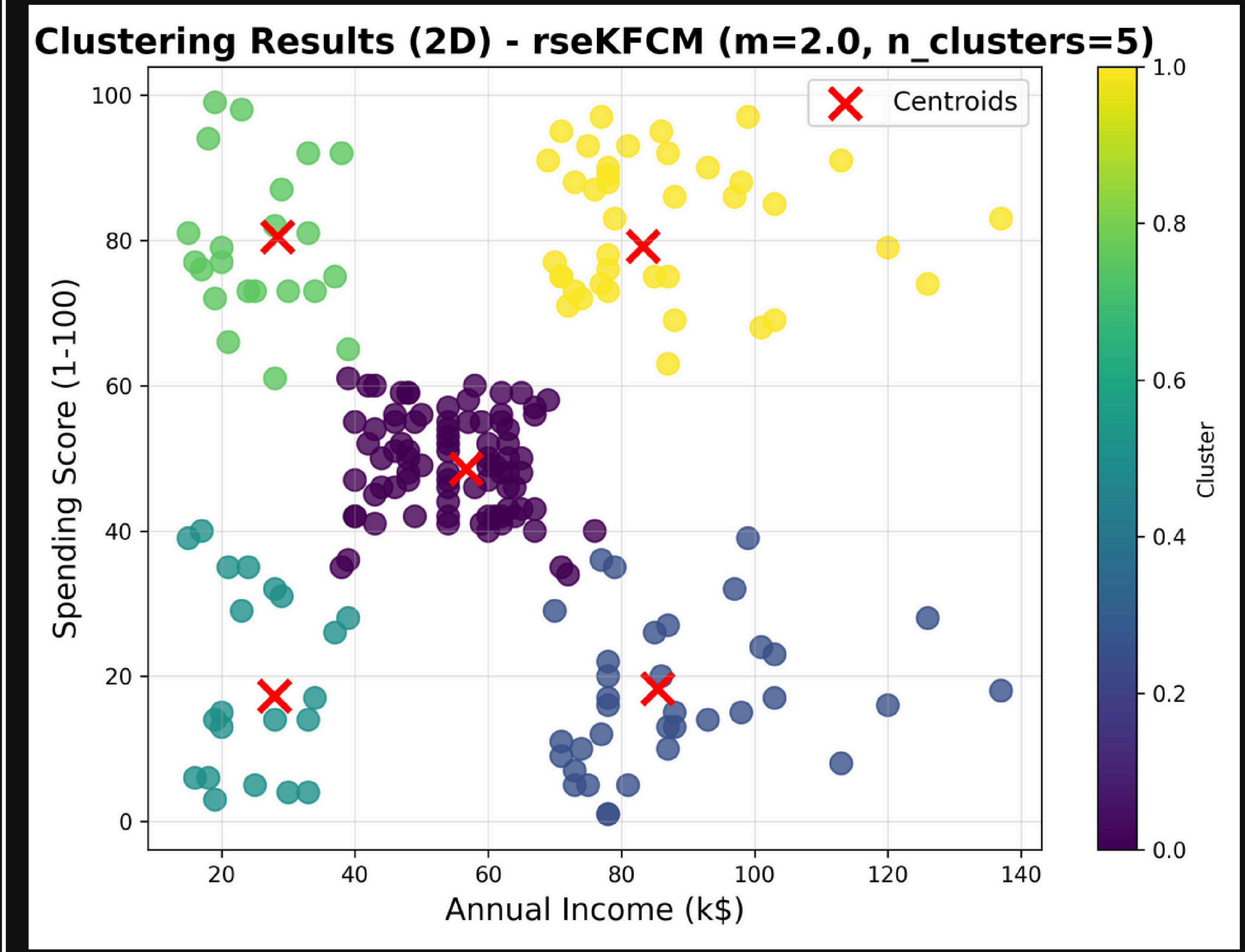
VISUALIZATION AND ANALYSIS

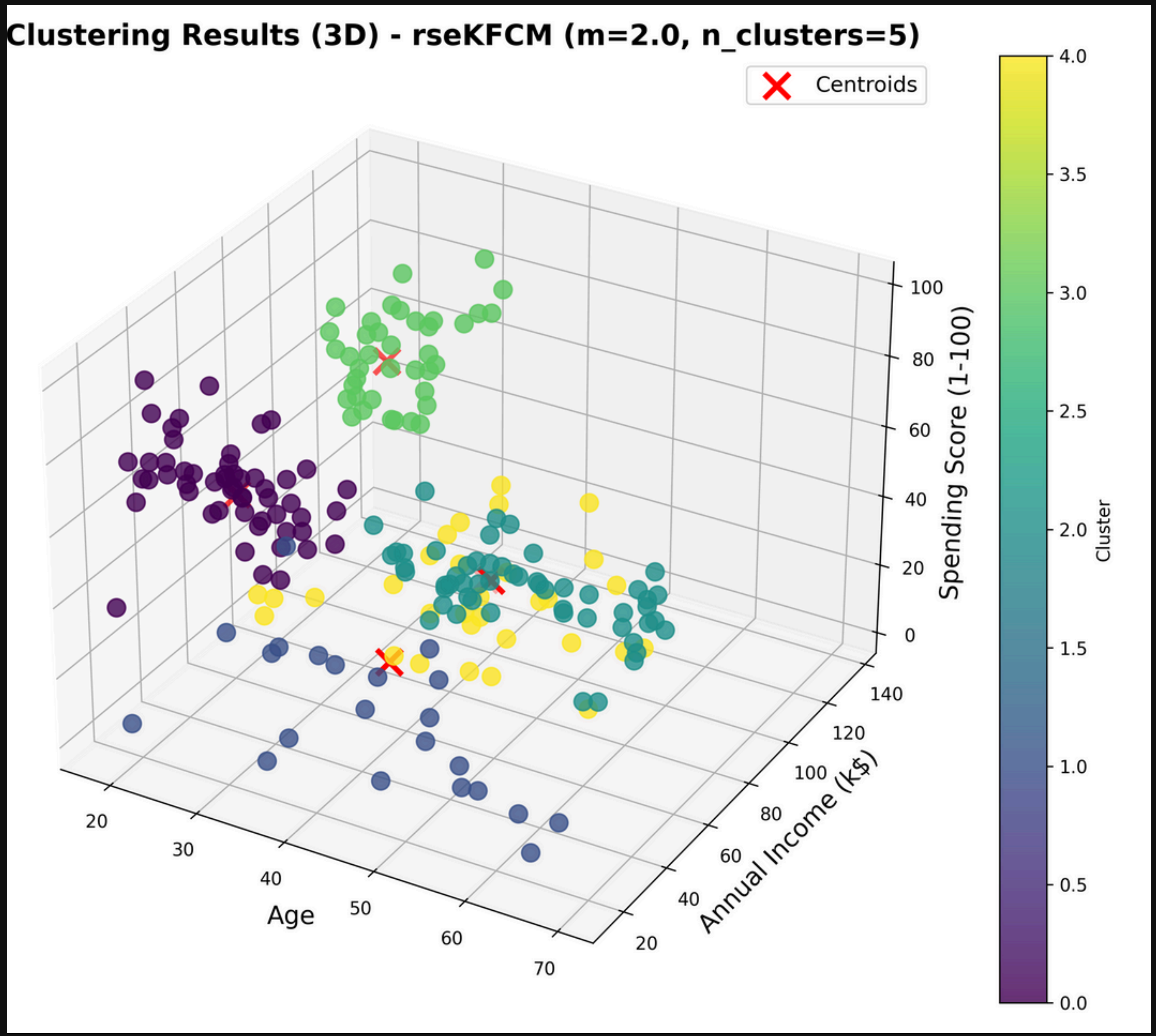


VISUALIZATION AND ANALYSIS

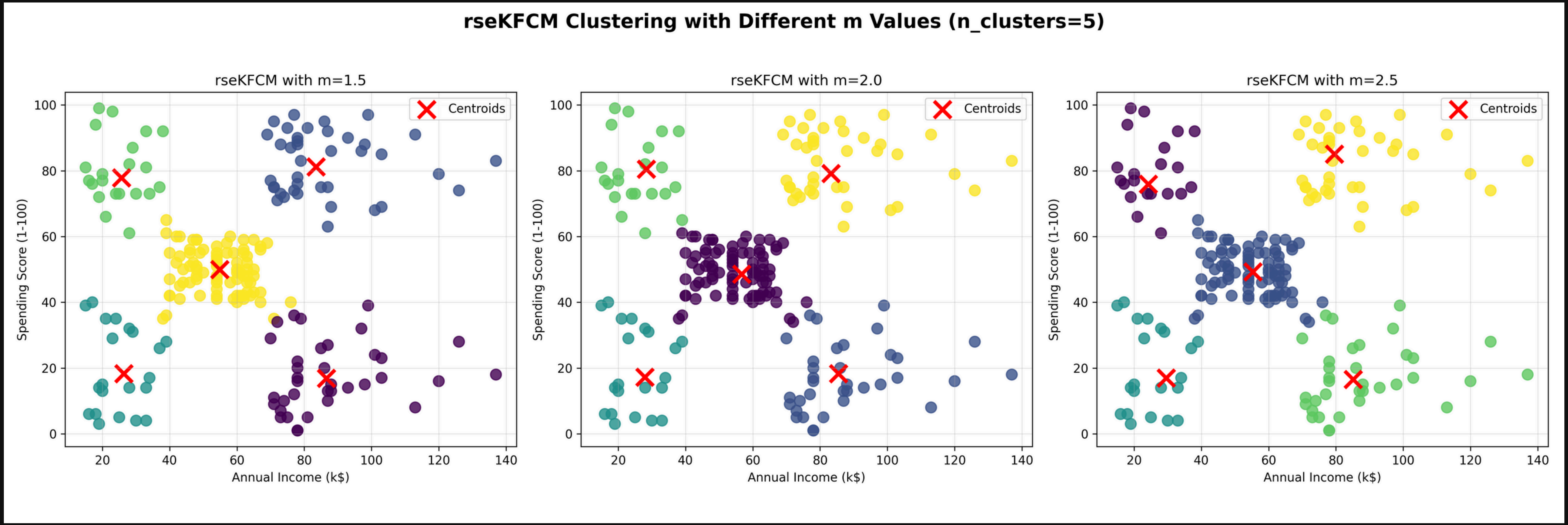


VISUALIZATION AND ANALYSIS

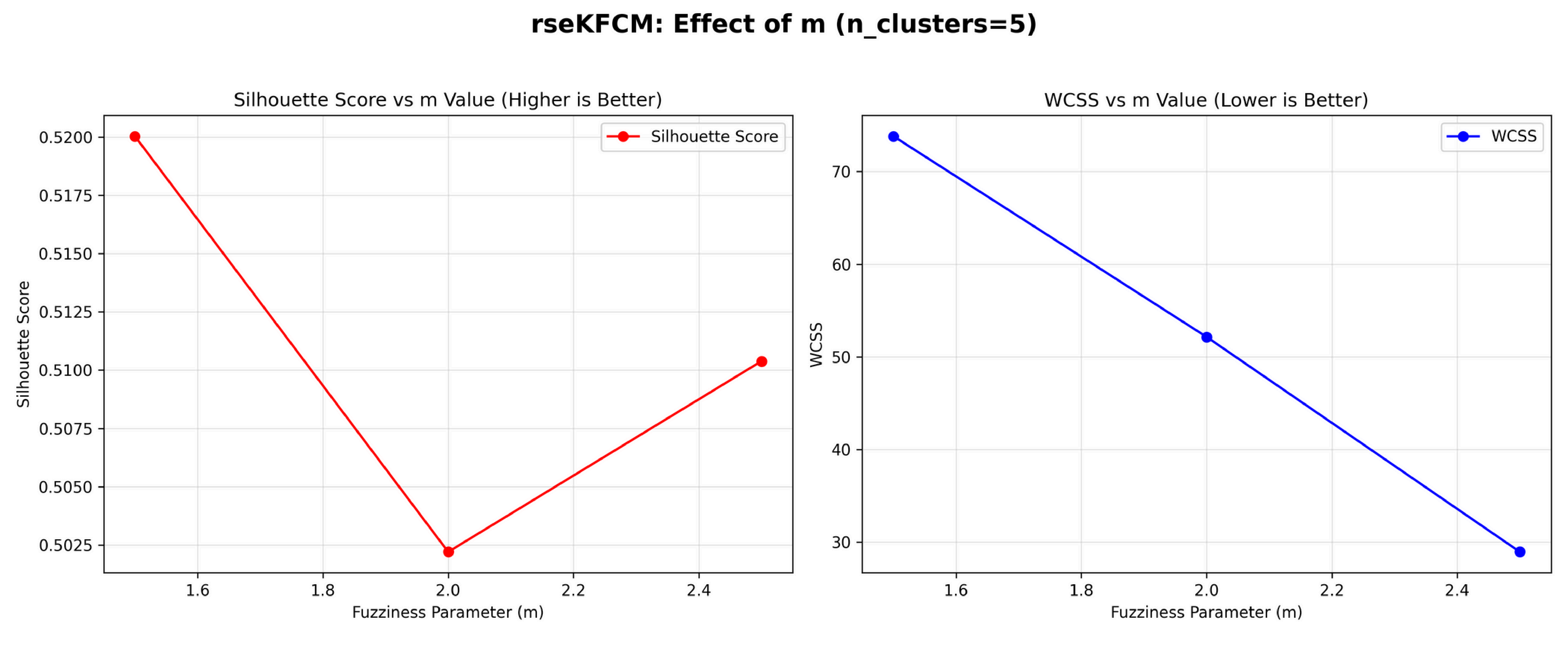




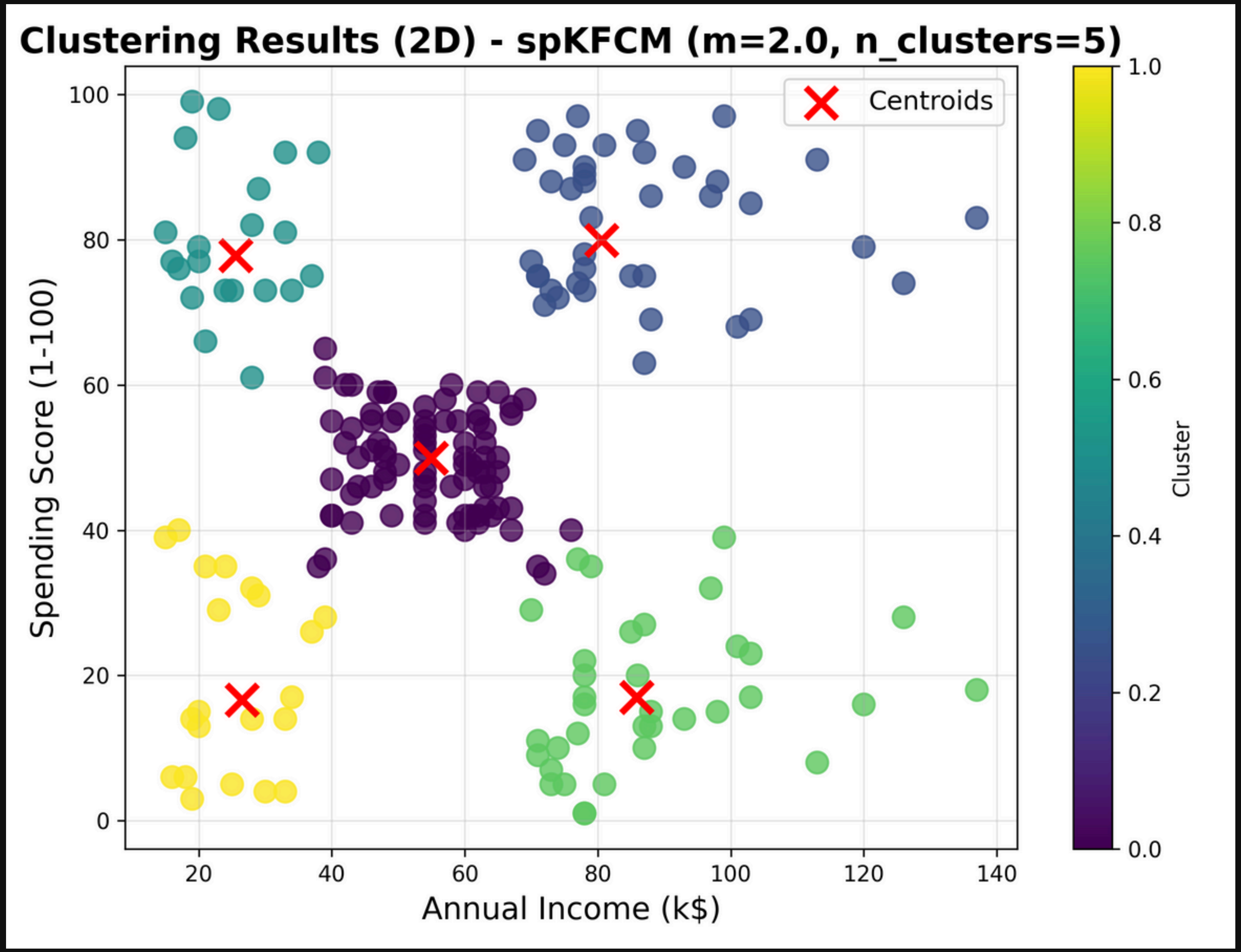
VISUALIZATION AND ANALYSIS

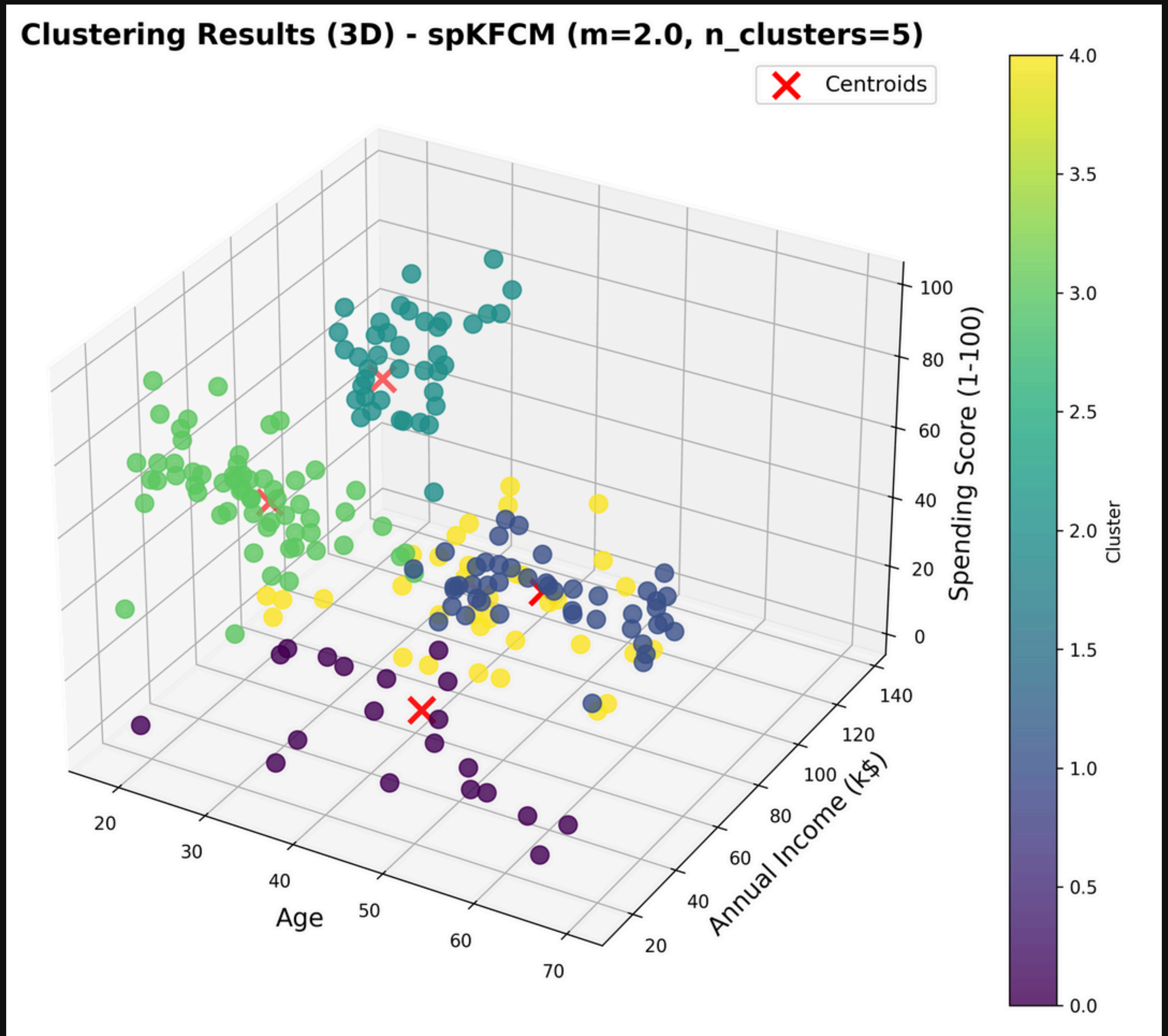


VISUALIZATION AND ANALYSIS

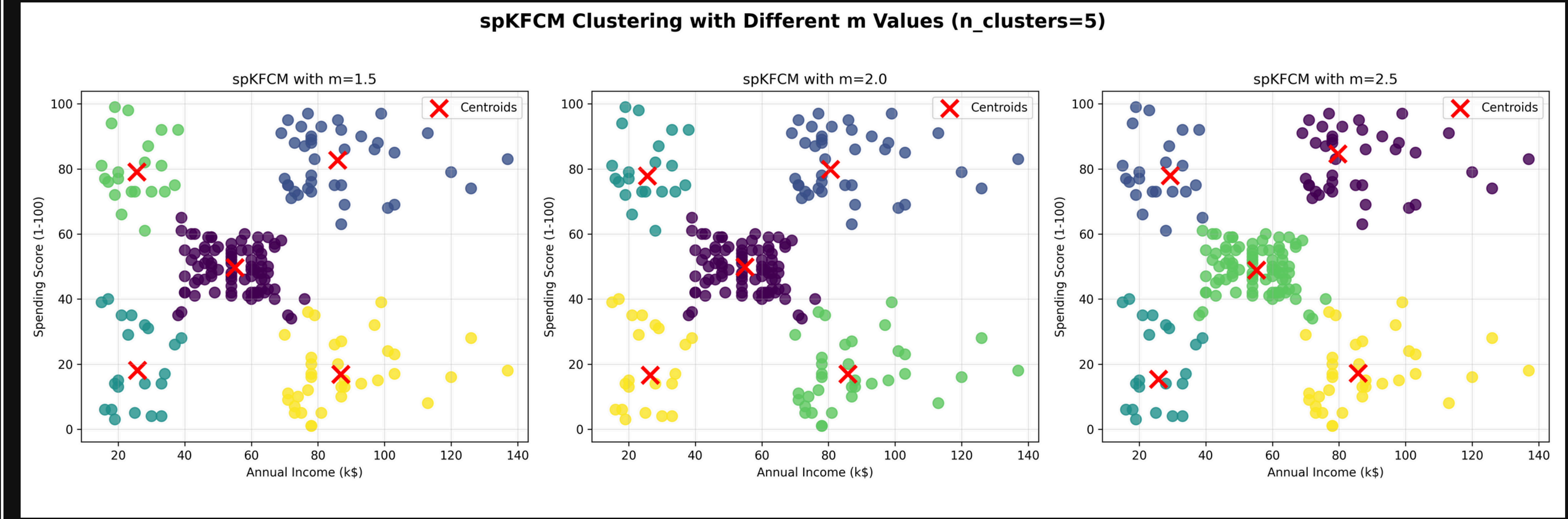


VISUALIZATION AND ANALYSIS

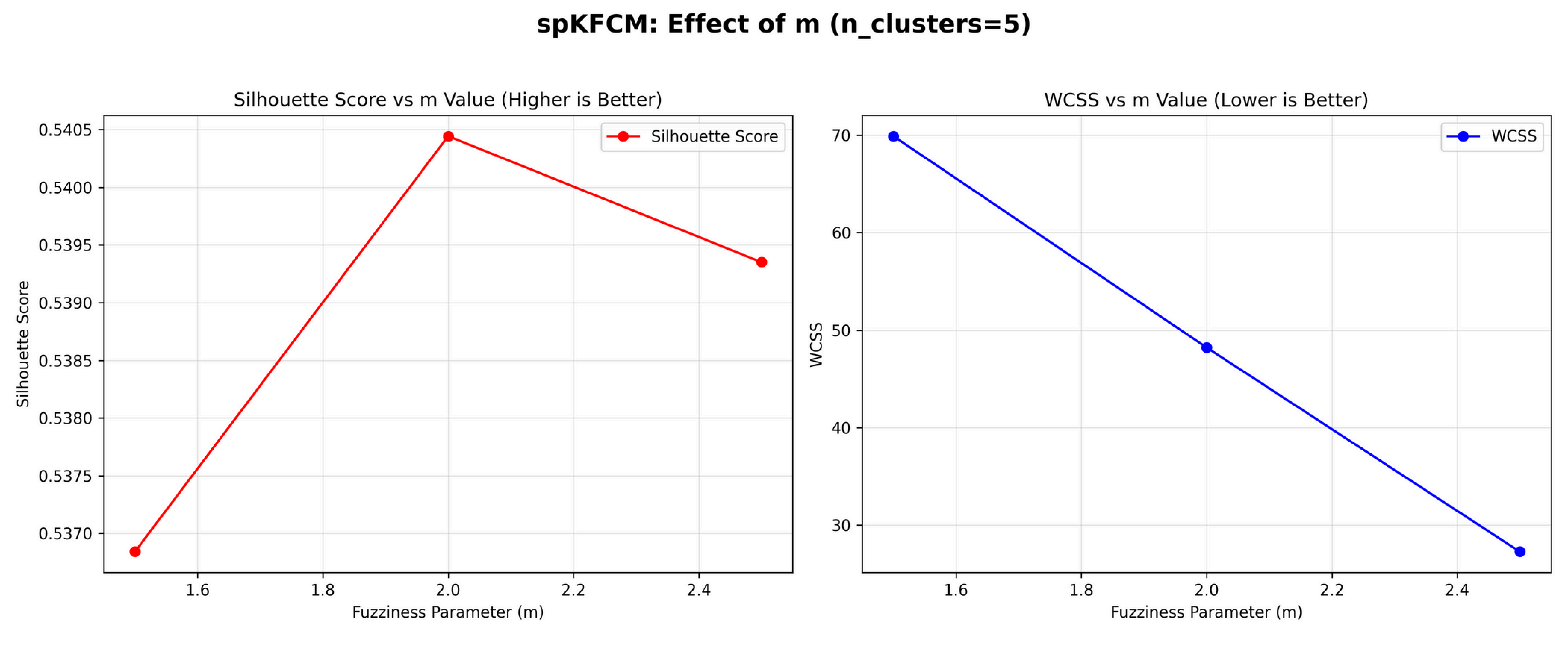




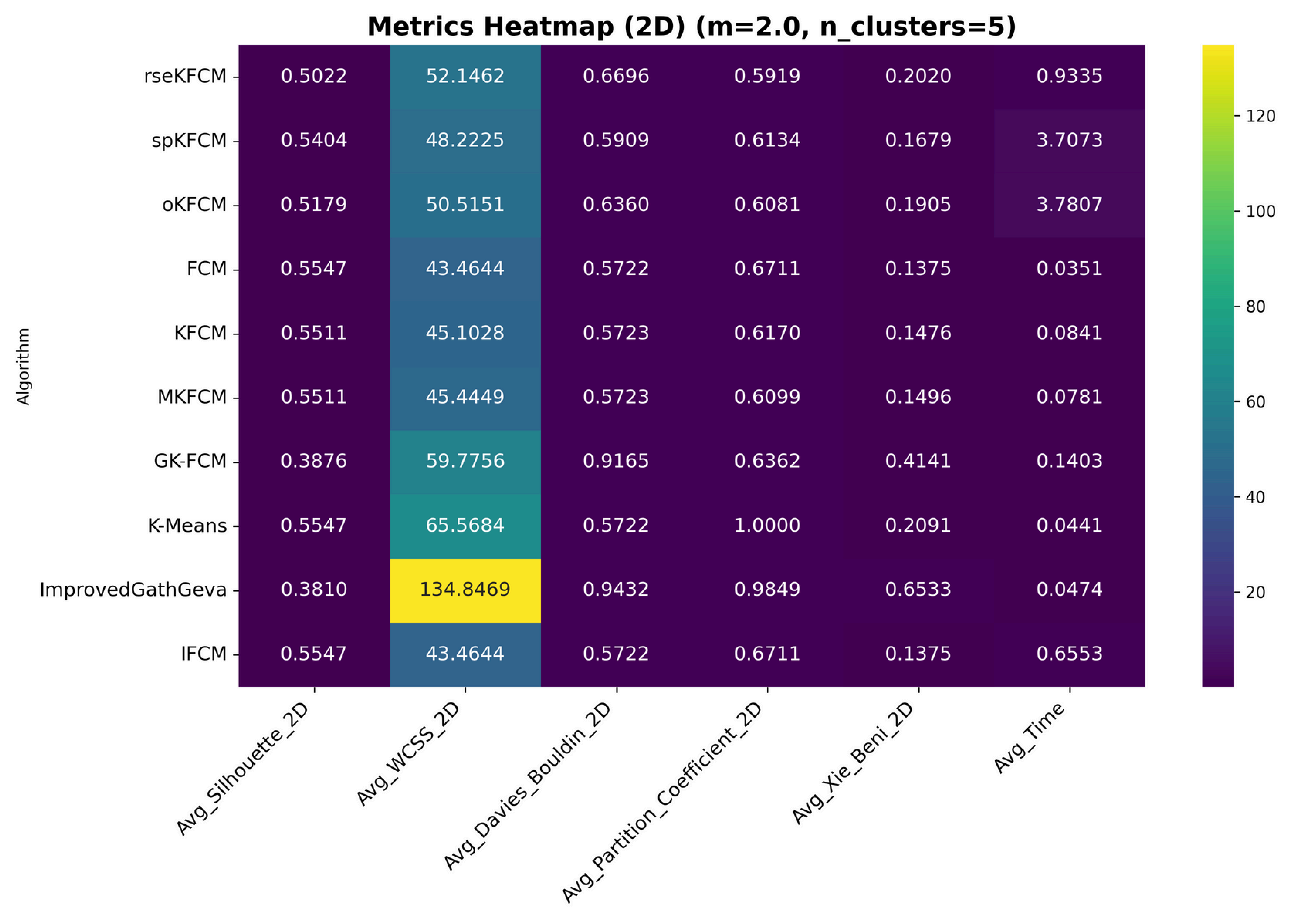
VISUALIZATION AND ANALYSIS



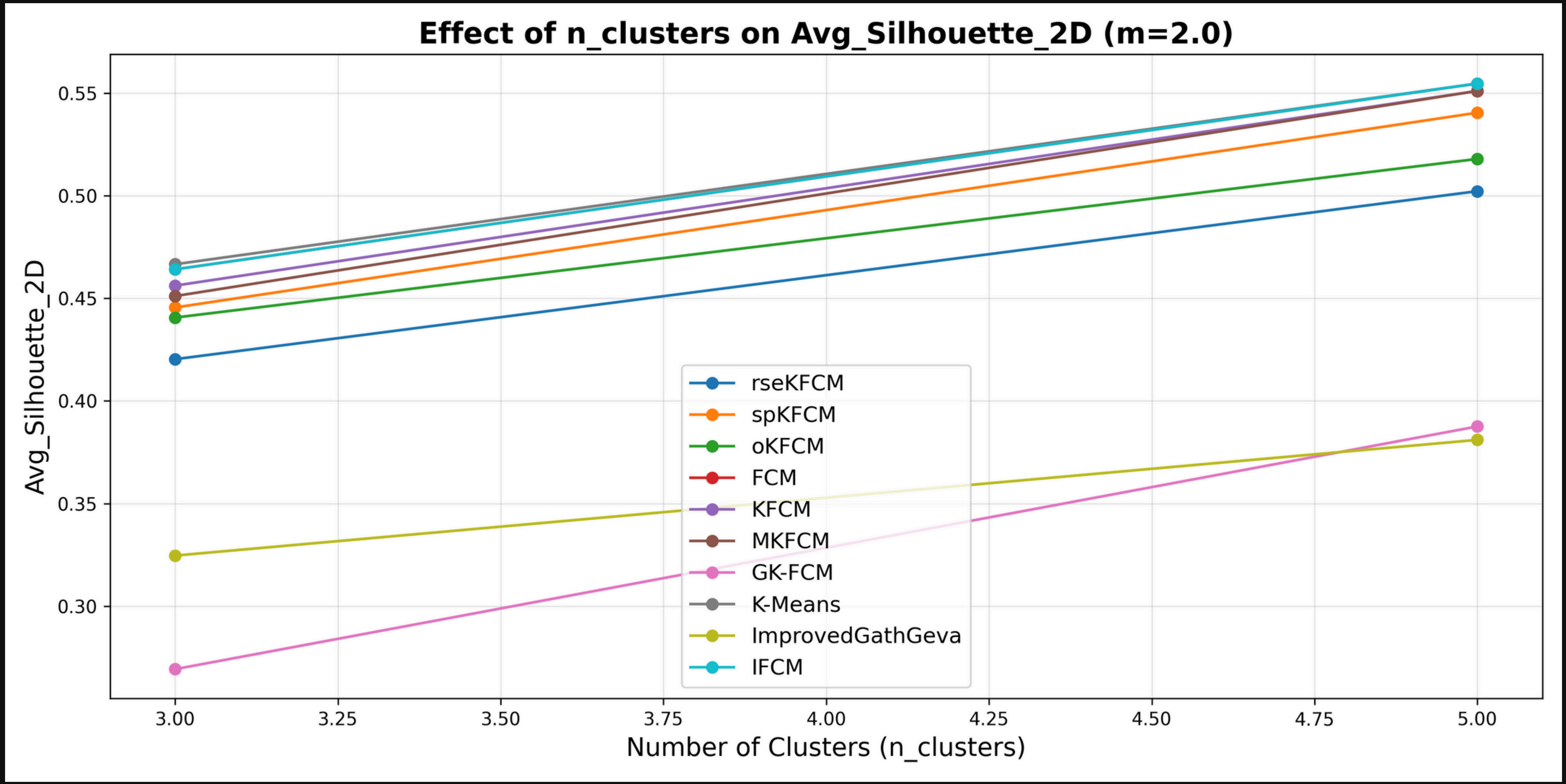
VISUALIZATION AND ANALYSIS



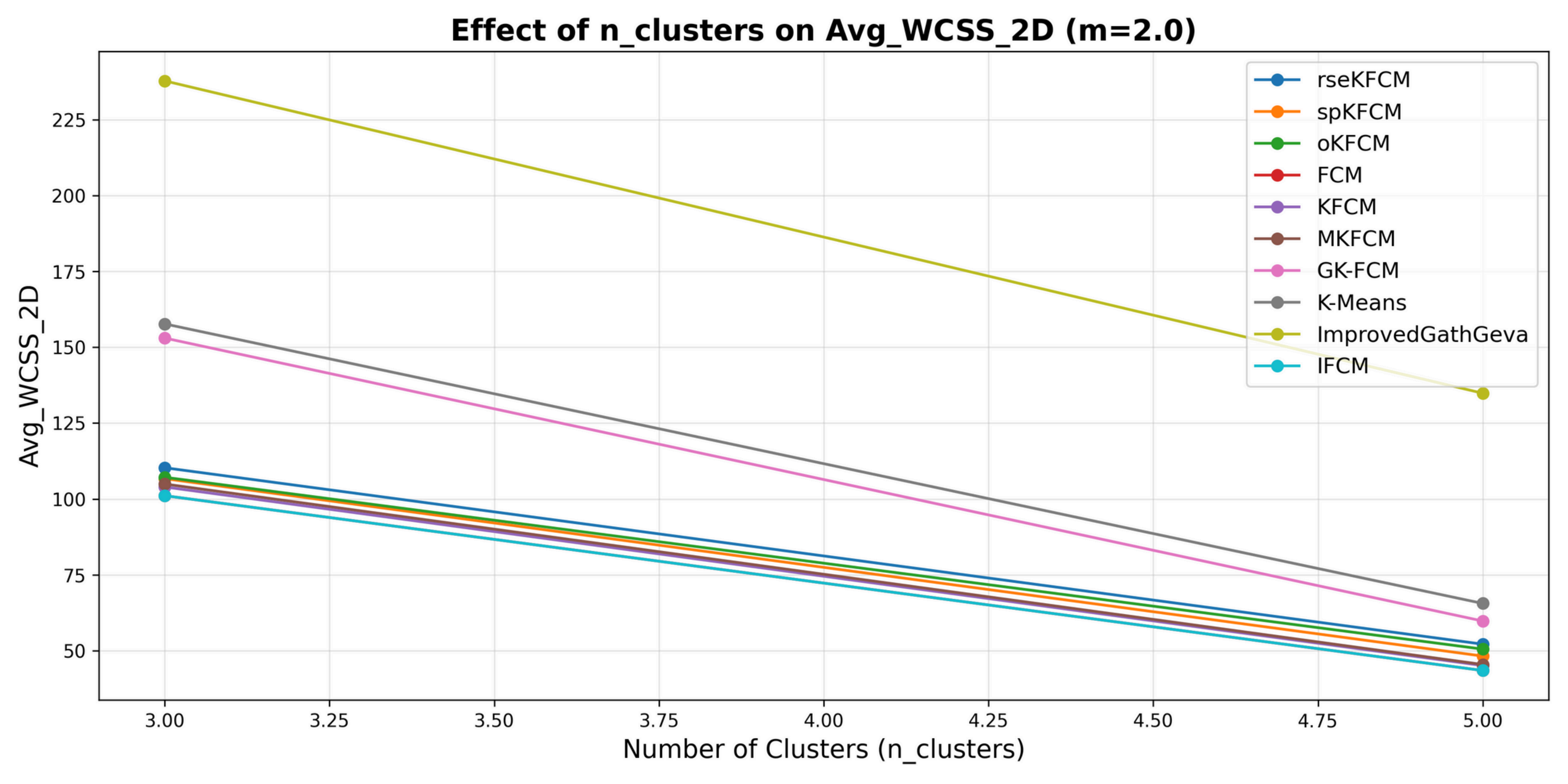
VISUALIZATION AND ANALYSIS



VISUALIZATION AND ANALYSIS



VISUALIZATION AND ANALYSIS



THANK YOU!

Have a great day ahead.