Interamerican University of Puerto Rico
Bayamon Campus
Department of Electrical and Computer Engineering

Project 1: Clustering

By

Edwin Soto Velázquez A00520488

COEN 4450
Data Science
Prof. Jaime Yeckle Sánchez

1) Introduction:

The purpose of this project is to apply concepts based on visualization, data management and grouping. Also, implementing in Python language, algorithms of grouping and index validation. All these concepts help strengthen current knowledge of Data Science and familiar topics. Finally, analyze the results of each dataset processed and view how clustering execution performed. This creates a fundament for developing new techniques when approaching related Data Science task.

Table 1: Task Completion

N	Description of Task	Completed Percentages (0-100%)
1	Dataset of Analysis and Graphs	100 %
2	Manual – Fuzzy C Means Algorithm	100 %
3	Automatic – Fuzzy C Means Algorithm	100 %
4	Implementation of index	100 %
Does the program run 100% of the tasks? (Yes / No)		Yes

2) Data set analysis:

Table 2: Analysis of Datasets

Dataset	Description	Groups
Datos_1	*a)* Group quantity: ***Based on observation we can determine the number of groups are 4*** *b)* Group size: ***The groups are based of cluster size 4*** *c)* Cohesion: ***There is cohesion between its groups since its points are separated from the inside of the group*** *d)* Separation: ***There is no separation between groups*** *e)* Density: ***There is density in 4 groups*** *f)* Shape: ***Oval circle shape of each group*** *g)* Noise: ***This dataset is not affected by noise***	4

Datos_2	*a)* Group quantity: ***Based on observation we can determine the number of groups are 4*** *b)* Group size: ***The groups are based of cluster size 4*** *c)* Cohesion: ***There is no cohesion between its groups since its points are close from the inside of the group*** *d)* Separation: ***There is slight separation between groups*** *e)* Density: ***There is density in 4 groups*** *f)* Shape: ***Round circle shape of each group*** *g)* Noise: ***This dataset is not affected by noise***	4
Datos_3	*a)* Group quantity: ***Based on observation we can determine the number of groups are 2*** *b)* Group size: ***The groups are based of cluster size 2*** *c)* Cohesion: ***There is no cohesion between its groups since its points are close from the inside of the group*** *d)* Separation: ***There is separation between groups*** *e)* Density: ***There is density in 2 groups*** *f)* Shape: ***Round circle shape of each group*** *g)* Noise: ***This dataset is not affected by noise***	2
Datos_4	*a)* Group quantity: ***Based on observation we can determine the number of groups are 1*** *b)* Group size: ***The groups are based of cluster size 1***	1

	c) Cohesion: ***There is cohesion between its groups since its points are separated from the inside of the group*** *d)* Separation: ***There is separation between points*** *e)* Density: ***There is density in 1 group*** *f)* Shape: ***Round circle shape of the group*** *g)* Noise: ***This dataset is affected by noise***	
Datos_5	*a)* Group quantity: ***Based on observation we can determine the number of groups are 3*** *b)* Group size: ***The groups are based of cluster size 3*** *c)* Cohesion: ***There is a slight cohesion between its groups since its points are separated from the inside of the group*** *d)* Separation: ***There is separation between groups*** *e)* Density: ***There is density in 3 groups*** *f)* Shape: ***Distorted round circle shape of each group*** *g)* Noise: ***This dataset is not affected by noise***	3
Datos_6	*a)* Group quantity: ***Based on observation we can determine the number of groups are 3*** *b)* Group size: ***The groups are based of cluster size 3*** *c)* Cohesion: ***There is cohesion between its groups since its points are separated from the inside of the group*** ***d)*** Separation: ***There is separation between groups*** *e)* Density: ***There is density in 3 groups***	3

	f) Shape: ***Distorted round circle shape of each group*** *g)* Noise: ***This dataset is affected by noise***	
Datos_7	*a)* Group quantity: ***Based on observation we can determine the number of groups are 2*** *b)* Group size: ***The groups are based of cluster size 2*** *c)* Cohesion: ***There is cohesion between its groups since its points are separated from the inside of the group*** *d)* Separation: ***There is separation between groups*** *e)* Density: ***There is density in 2 groups*** *f)* Shape: ***Oval circle shape of each group*** *g)* Noise: ***This dataset is not affected by noise***	2
Datos_8	*a)* Group quantity: ***Based on observation we can determine the number of groups are 2*** *b)* Group size: ***The groups are based of cluster size 2*** *c)* Cohesion: ***There is cohesion between its groups since its points are separated from the inside of the group*** *d)* Separation: ***There is separation between groups*** *e)* Density: ***There is density in 2 groups*** *f)* Shape: ***Round circle shape of each group*** *g)* Noise: ***This dataset is not affected by noise***	2
Datos_9	*a)* Group quantity: ***Based on observation we can determine the number of groups are 1*** *b)* Group size: ***The groups are based of cluster size 1***	1

	c) Cohesion: ***There is cohesion between its groups since its points are separated from the inside of the group*** *d)* Separation: ***There is no separation between groups*** *e)* Density: ***There is density in 1 group*** *f)* Shape: ***Distorted round circle shape of each group*** *g)* Noise: ***This dataset is affected by noise***	
Datos_10	*a)* Group quantity: ***Based on observation we can determine the number of groups are 2*** *b)* Group size: ***The groups are based of cluster size 2*** *c)* Cohesion: ***There is cohesion between its groups since its points are separated from the inside of the group*** *d)* Separation: ***There is slight separation between groups*** *e)* Density: ***There is density in 2 groups*** *f)* Shape: ***Oval circle shape of each group*** *g)* Noise: ***This dataset is not affected by noise***	2
Datos_11	*a)* Group quantity: ***Based on observation we can determine the number of groups are 2*** *b)* Group size: ***The groups are based of cluster size 2*** *c)* Cohesion: ***There is not much cohesion between its groups since its points are close from the inside of the group*** *d)* Separation: ***There is separation between groups*** *e)* Density: ***There is density in 2 groups***	2

	f) Shape: ***Round circle shape of each group*** *g)* Noise: ***This dataset is not affected by noise***	
Datos_12	*a)* Group quantity*: **Based on observation we can determine the number of groups are 2*** *b)* Group size: ***The groups are based of cluster size 2*** *c)* Cohesion: ***There is cohesion between its groups since its points are separated from the inside of the group*** *d)* Separation: ***There is separation between groups*** *e)* Density: ***There is density in 2 groups*** *f)* Shape: ***Distorted round circle shape of each group*** *g)* Noise: ***This dataset is affected by noise***	2

3) Results of Fuzzy C Means:

a) *Best value of M:*

Based on testing every possible option for M, the best overall M value is 2. It has more accurate results when detecting groups.

b) *Distance metrics:*

Although, many distance metrics are good in certain datasets, the average distance metric that had more accurate results is the Euclidean metric. It marks the closest distance between each point.

c) *Initialization methods:*

The best initialization method is with random numbers, since it evaluates with more accuracy.

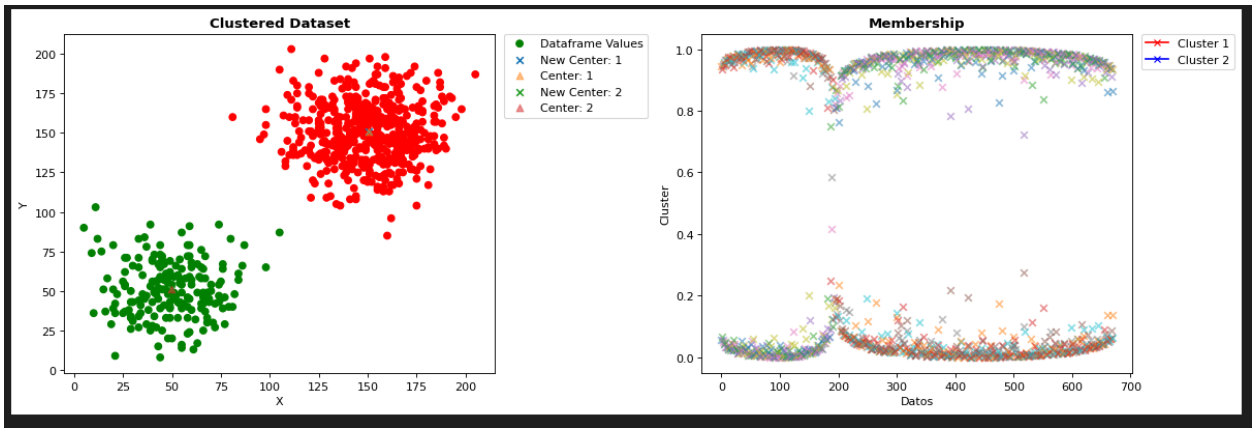
d) *The effect of noise:*

The effect of noise is in some datasets since accuracy loss increases when detecting groups. It is unclear when calculating the membership since the effect of noise disorientate some values.

e) *Detection of efficient groups:*

If you observe the following graph, we can see how efficient it detects groups and the metrics selected (*Cluster size: 2, Euclidean distance metric, M value: 2, Random Elements Initialization*) demonstrate efficiency when detecting groups.

Graph 1: Dataset 11



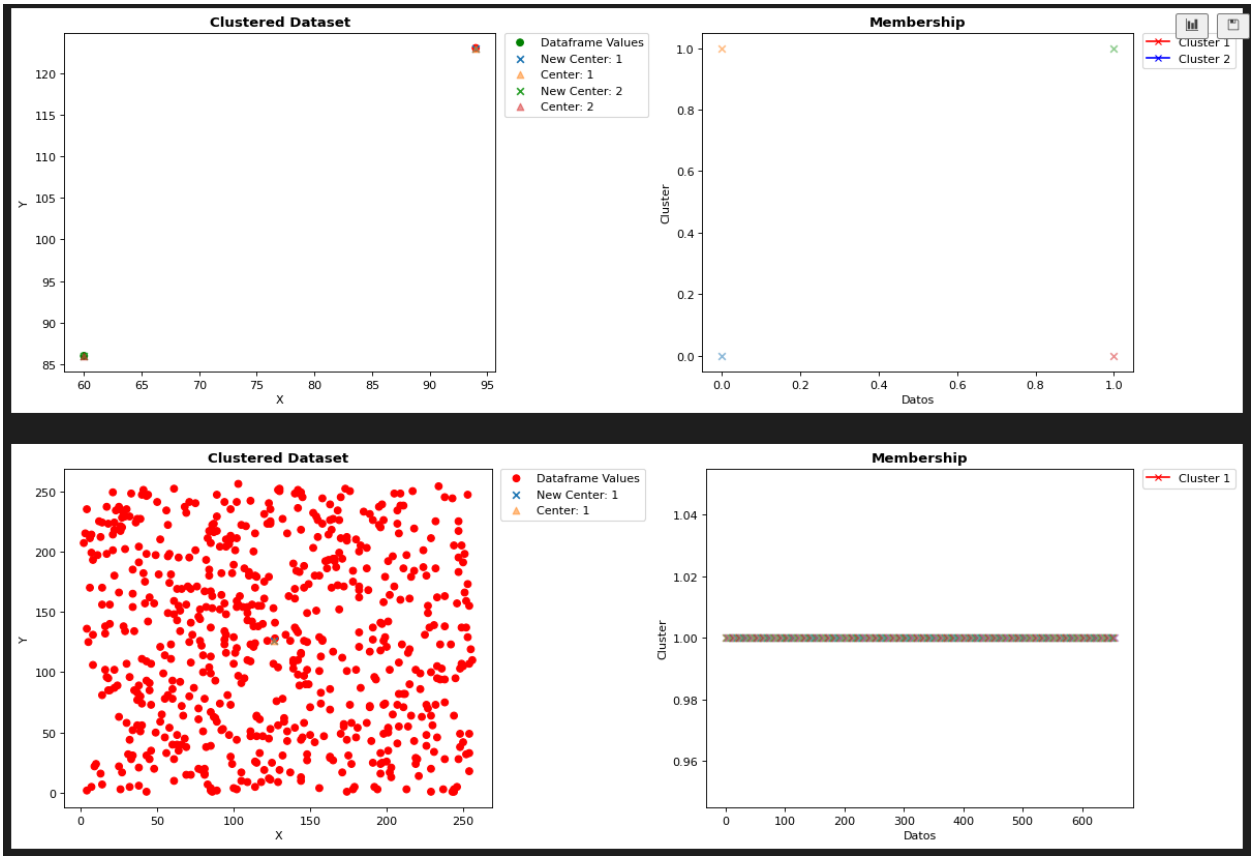
f) Which parameters are efficient?

The most efficient parameters are: Euclidean distance metric, M value 2, Random Elements Initialization.

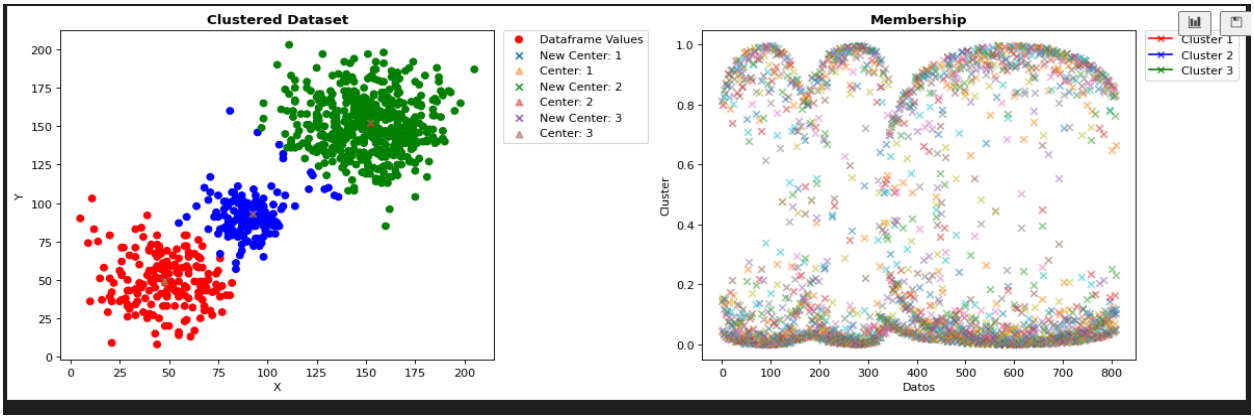
g) Detection of groups from datasets with different: size, density, forms, and sound:

As we can see in the following graphs, you can stand firm that the algorithm is performing an excellent job at clustering different sizes of groups based on their density and form even if noise is affecting some of the datasets.

Graph 2: Dataset 3 (Superior), Dataset 4 (Inferior)



Graph 3: Dataset 5



Important: *Evidence that support these arguments are in the Appendix section (Part 2: Graphs) containing all the Graphs with every possible option tested.*

4) Implementation of indexes:

The strategy I used to evaluate all the index was verifying each value from the dataset pertains to its respective groups to achieve the most accurate results as possible. Then I compared the index returned value and compared to my proposed value for that specific dataset. As a result, if we analyze all the values compared to the one, I proposed we can see over a 78 % percent match. Having the rest of the 21 % in a miss of calculation, knowing the indexes can miss since there not perfect. For example, my proposed cluster size for Dataset 1 is 4; PC tells us that the number that was proposed is really close to what we proposed, and FS confirms our proposed value. However, the Ball index proposed a far-off number than what I suspected. Finally, the following table demonstrates the results of the indexes for each dataset with the best possible metrics.

Table 3: Results of index

Dataset	**Number of Groups**			
	Proposed	*PC*	*FS*	*Ball*
Datos_1	4	1	4	2
Datos_2	4	1	3	5
Datos_3	2	1	2	2
Datos_4	1	1	3	1
Datos_5	3	1	3	3
Datos_6	3	1	3	4
Datos_7	2	1	2	2
Datos_8	2	1	3	2
Datos_9	1	1	1	1
Datos_10	2	1	2	2

Datos_11	2	1	2	2
Datos_12	2	1	2	3

5) Conclusion:

The validation of data and correct interpretation can deliver a powerful solution to many problems. This project provided me with a challenge where I could learn how clustering works in the Data Science field and to try to create my own algorithm to perform such technique. I learned how to analyze, create complex mathematical formulas into code and process different types of data being ingested by CSV formatted documents. For me, this was a great source of knowledge so that I can improve my skills in programming and other areas. It took many hours of dedication and research which made me more interested in solving each issue I encountered along the process. In the end, it was rewarding running the program a see how my algorithm accurately determined each group from the datasets, color them and tell me the results of its calculation accomplishing the goal to cluster each group.

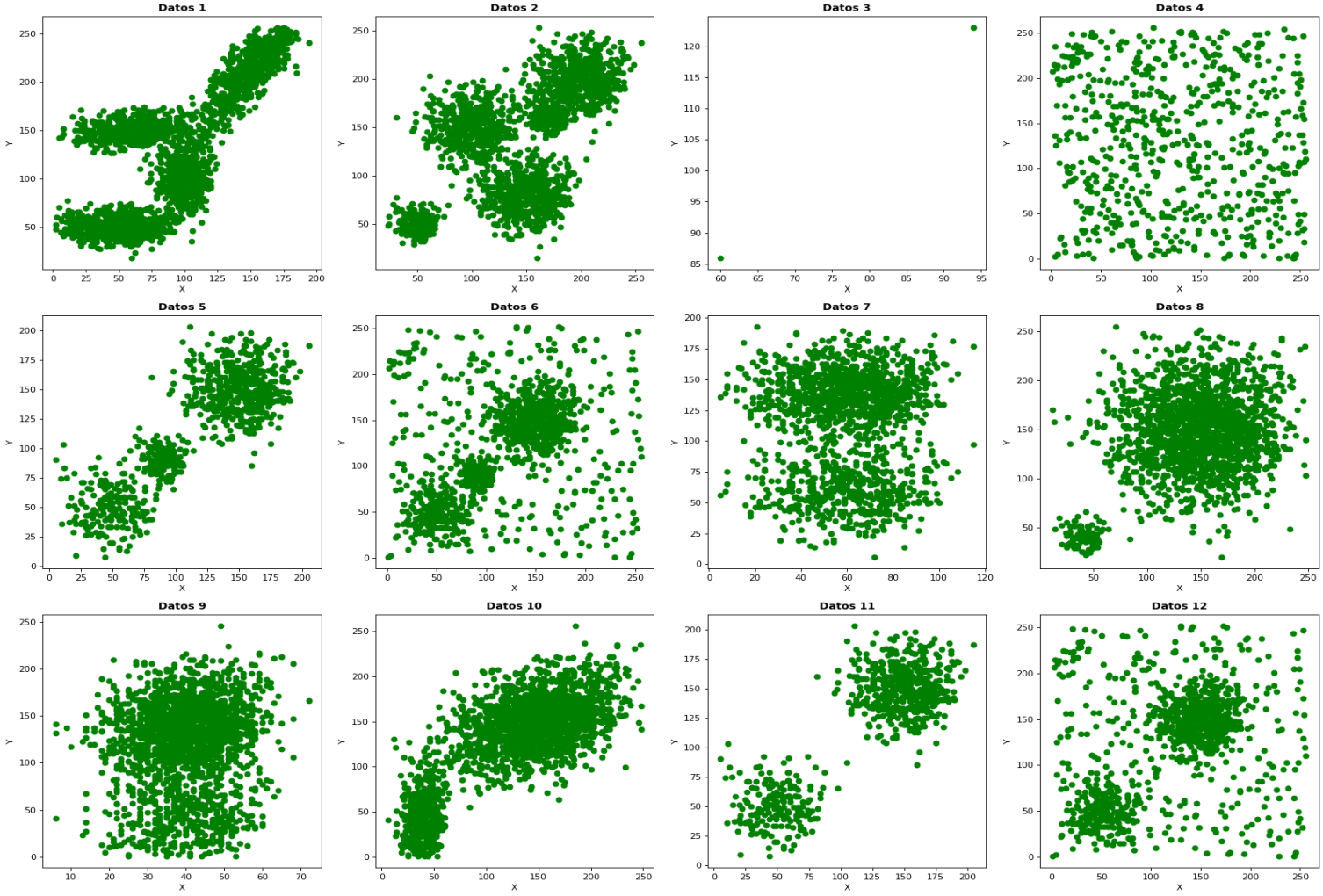
6) Appendix:

I. Table with best parameters for each dataset:

Dataset	Methods of Distance	Cluster Size	M	Initialization of centers
Datos_1	Euclidean	4	2	Random Elements
Datos_2	Chebyshev	4	2	First Elements
Datos_3	Manhattan	2	2	First Elements
Datos_4	Chebyshev	1	2	First Elements
Datos_5	Chebyshev	3	2	Random Elements
Datos_6	Euclidean	2	2	Random Elements
Datos_7	Euclidean	2	2	First Elements
Datos_8	Manhattan	1	2	Random Elements
Datos_9	Manhattan	1	2	First Elements
Datos_10	Manhattan	2	2	First Elements
Datos_11	Euclidean	2	2	Random Elements
Datos_12	Euclidean	2	2	

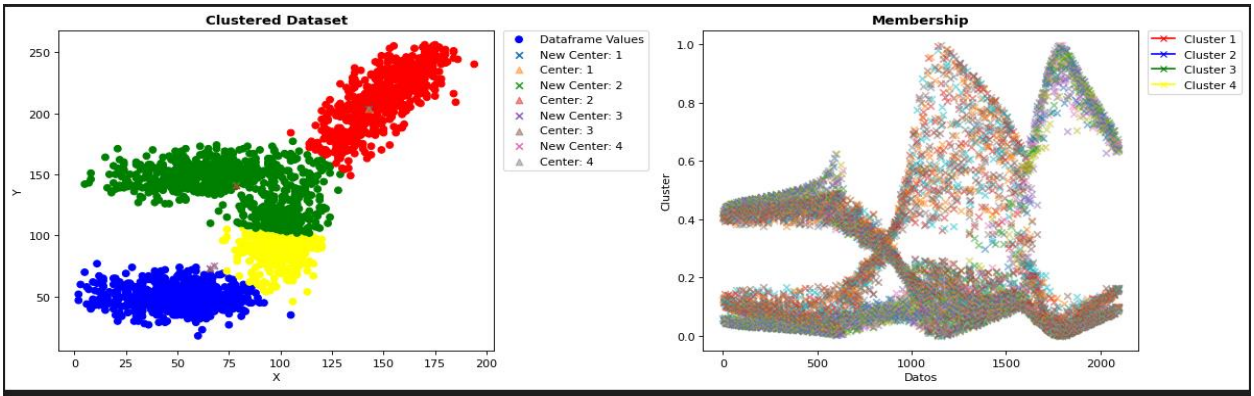
II. Graphs of all datasets:

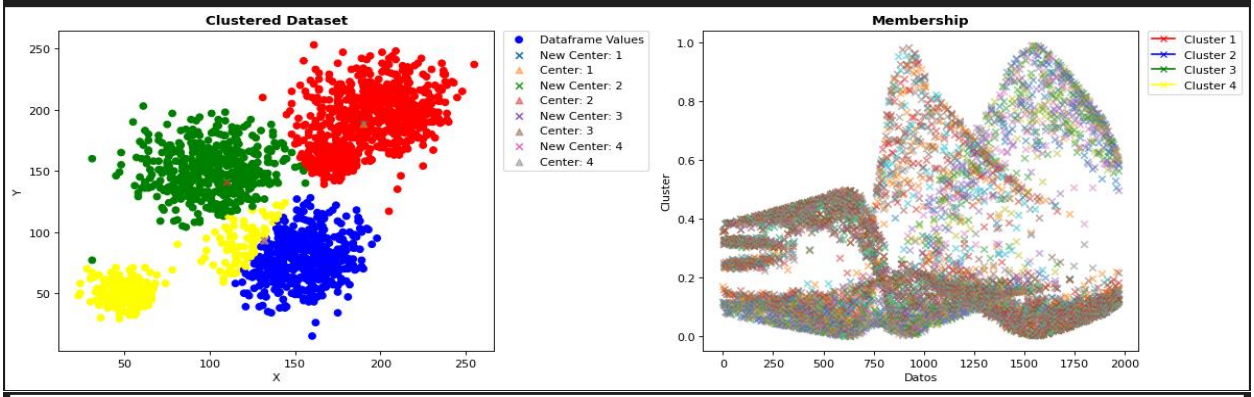
A. Raw Data

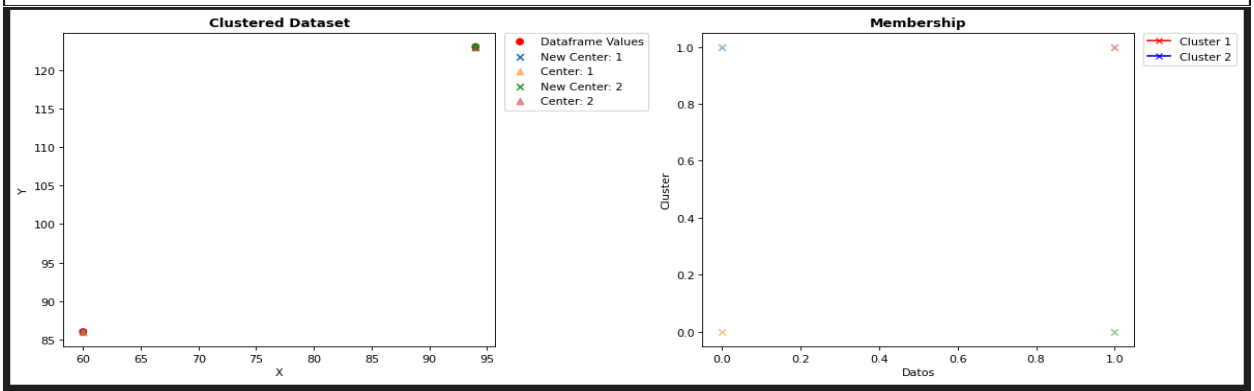


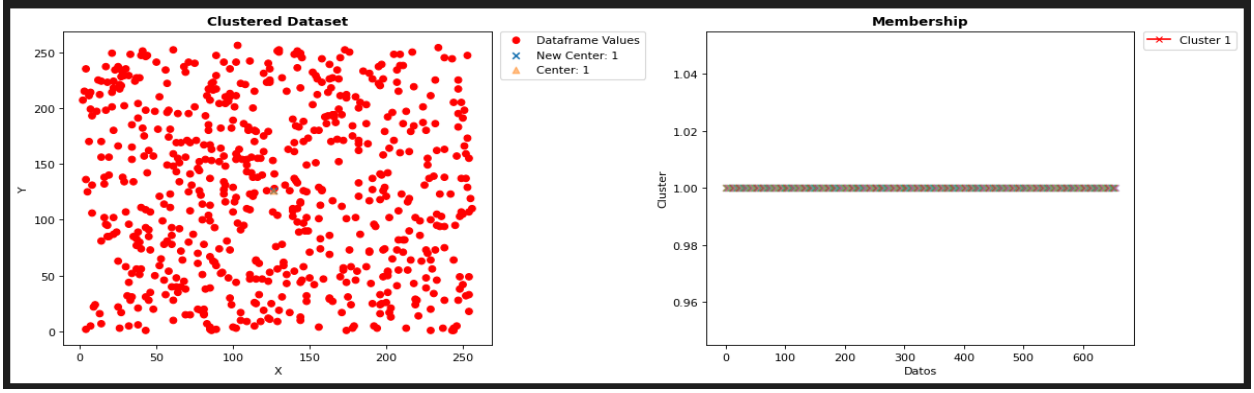
B. FCM

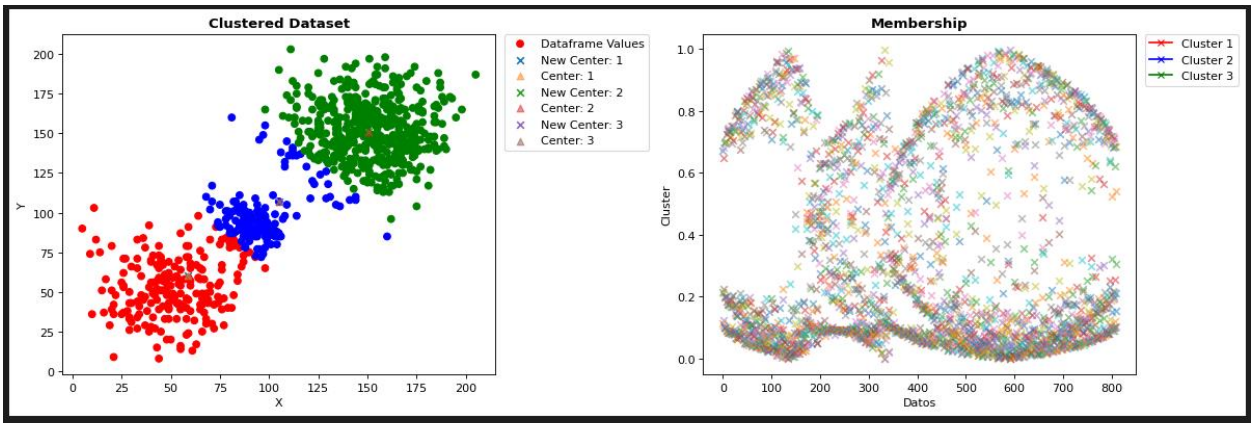
a. M = **1.5**; Initialization of Centers = **First Elements**; Distance Method: **Euclidean**

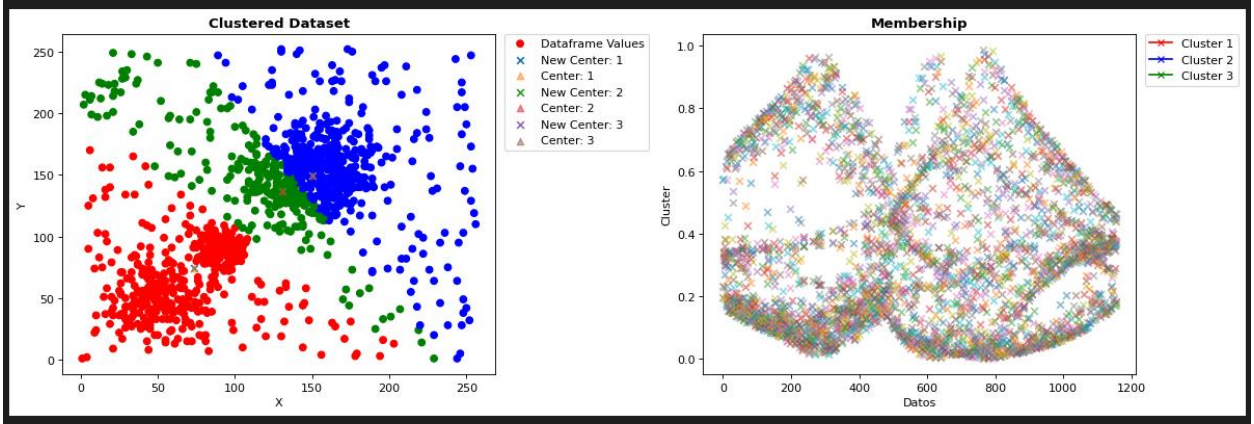


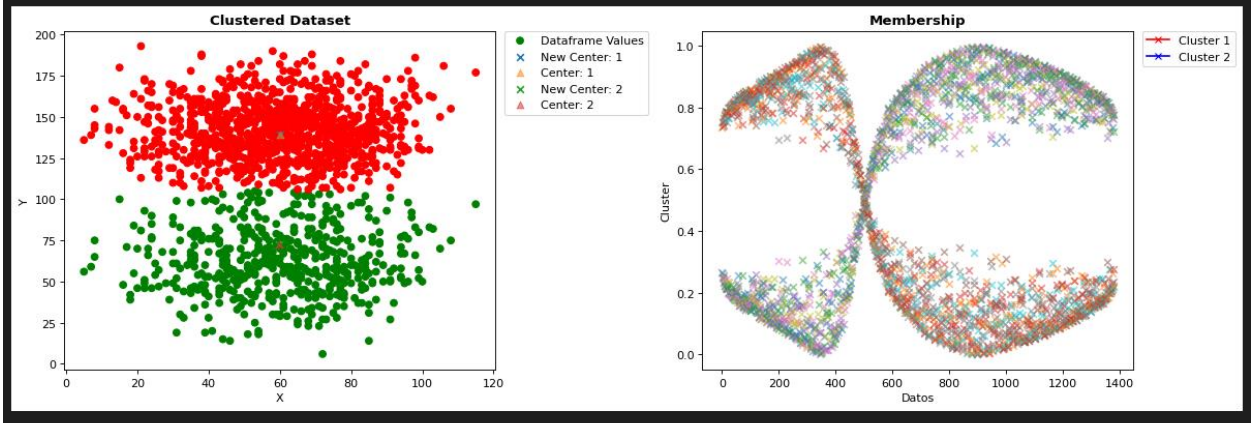


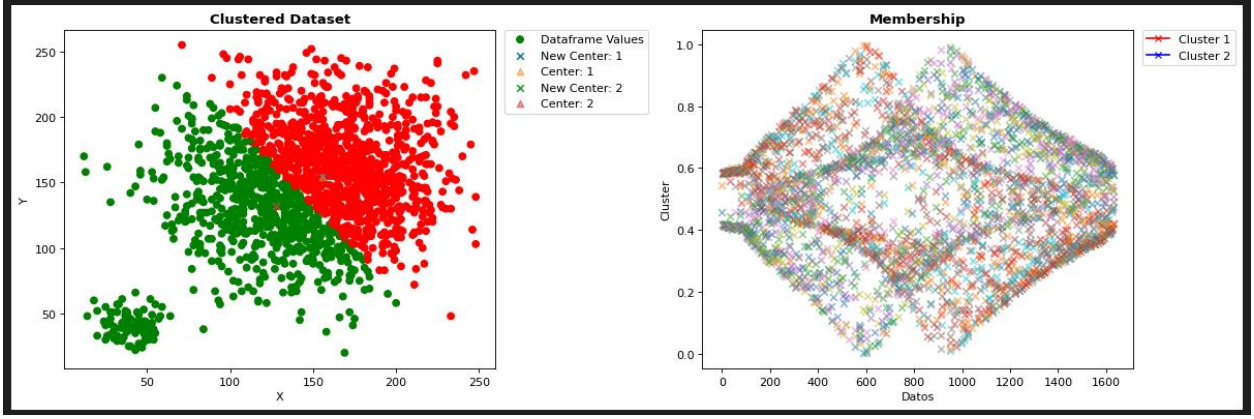


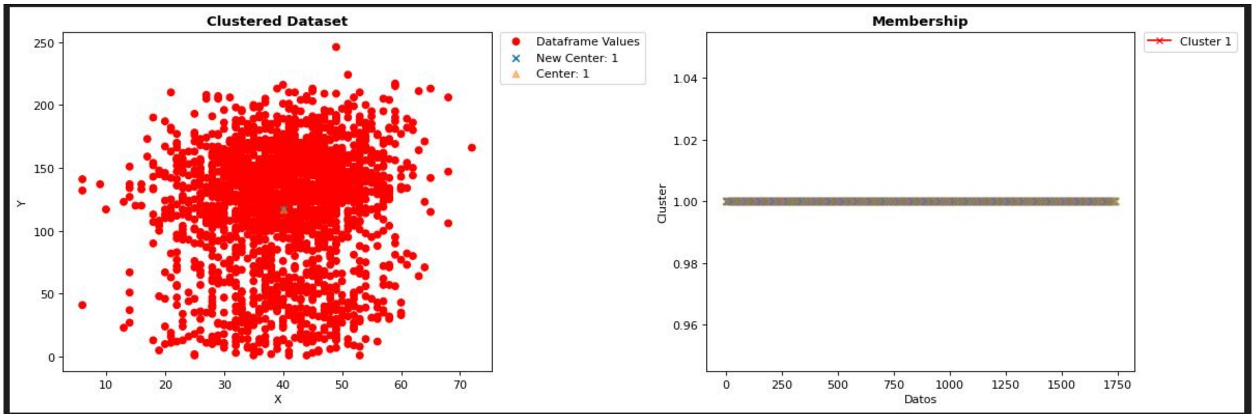


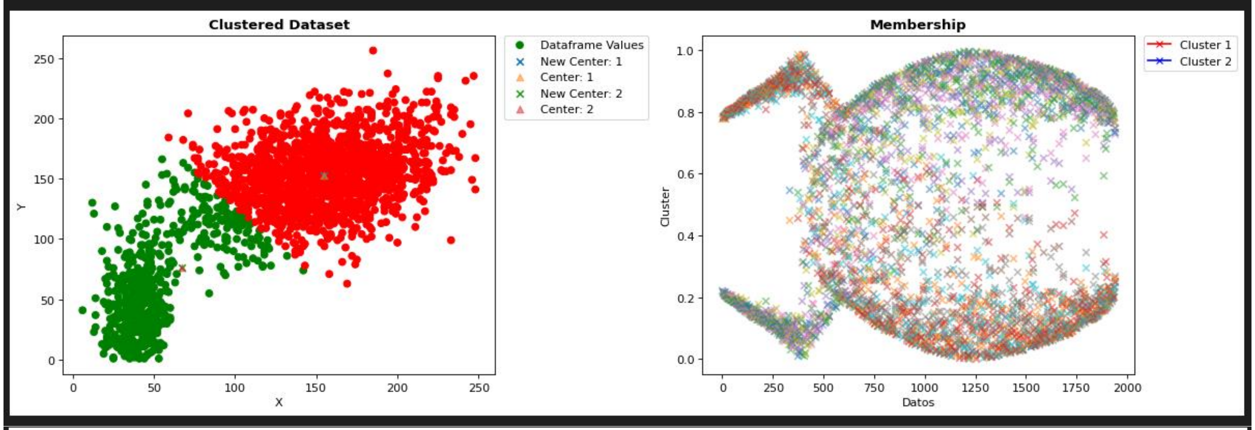

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

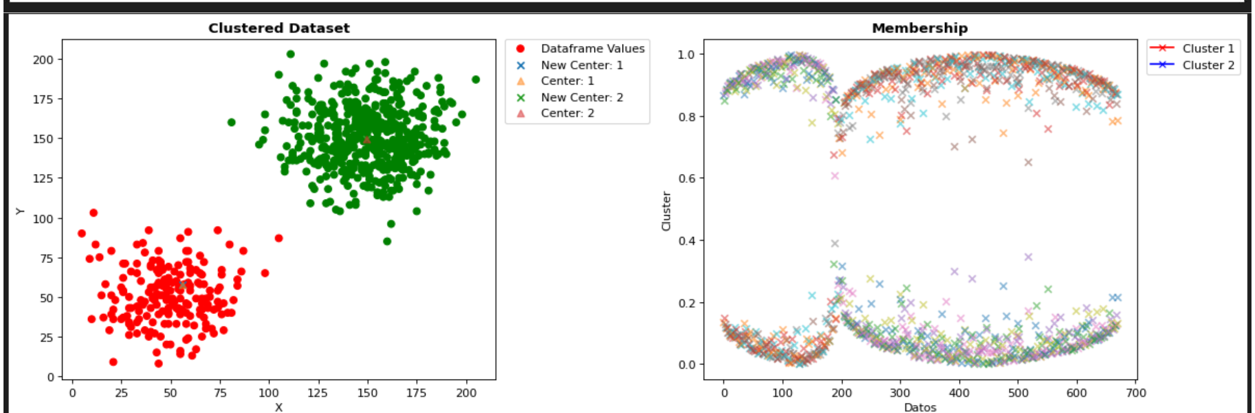

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

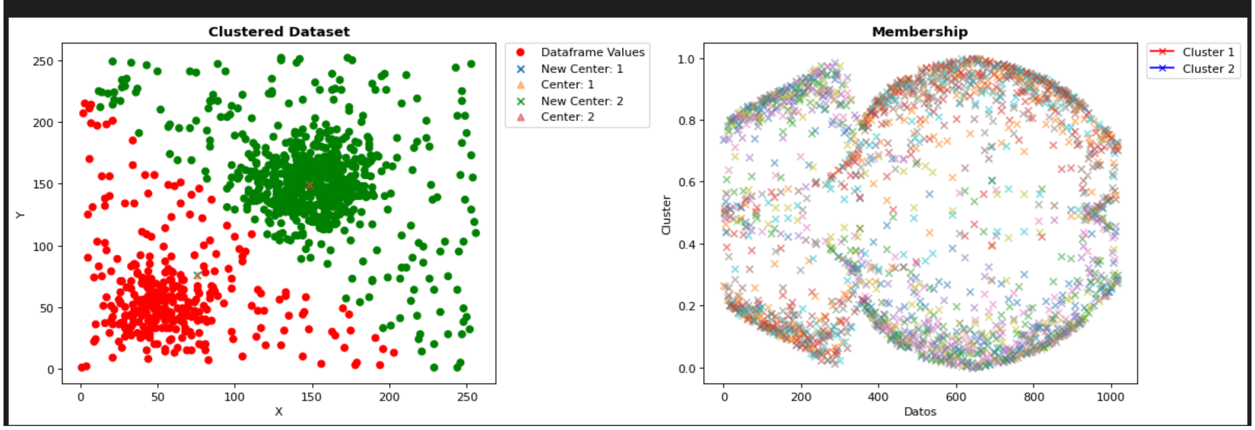

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos


Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

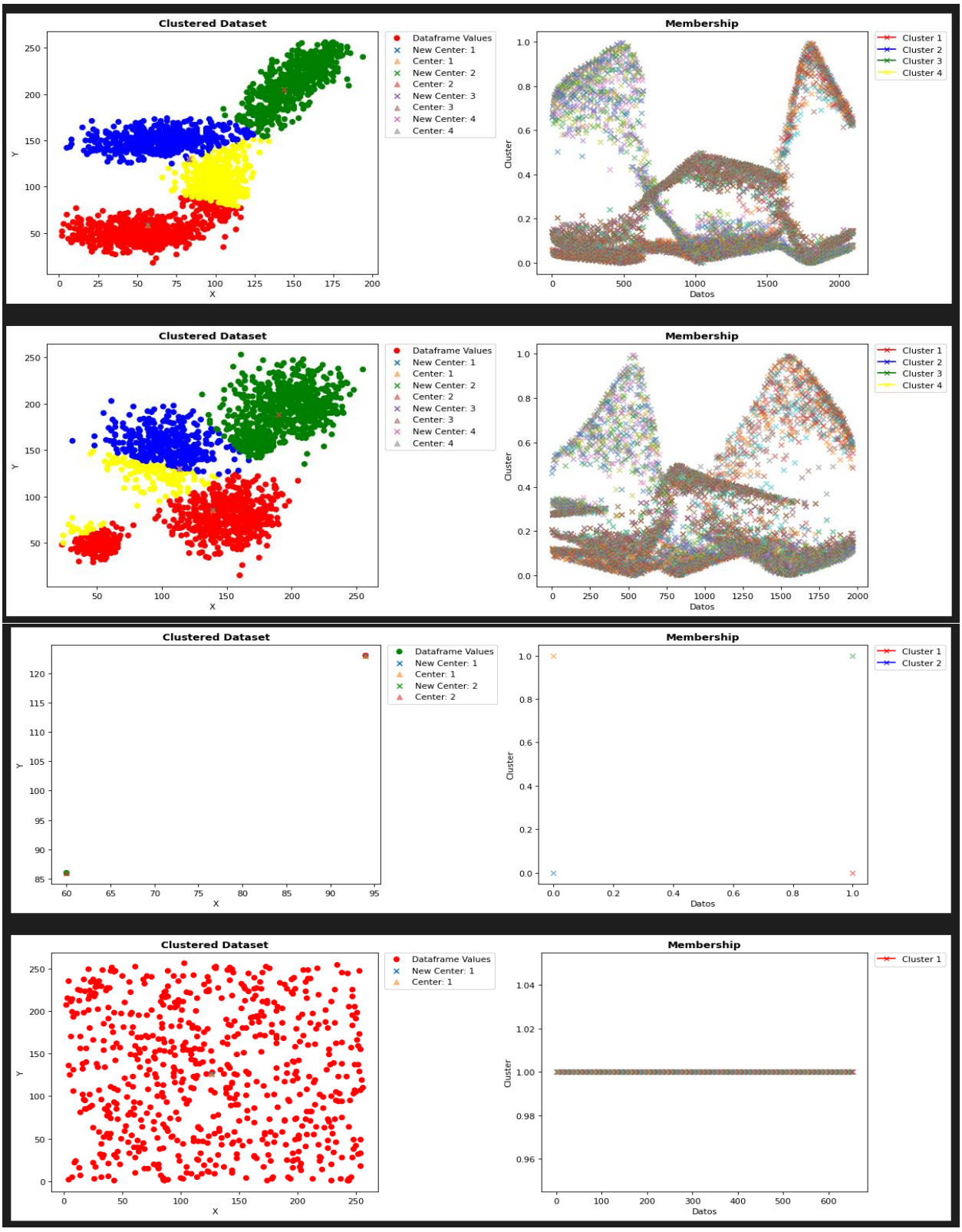

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
X
Y
Membership
Cluster 1
Datos
Cluster

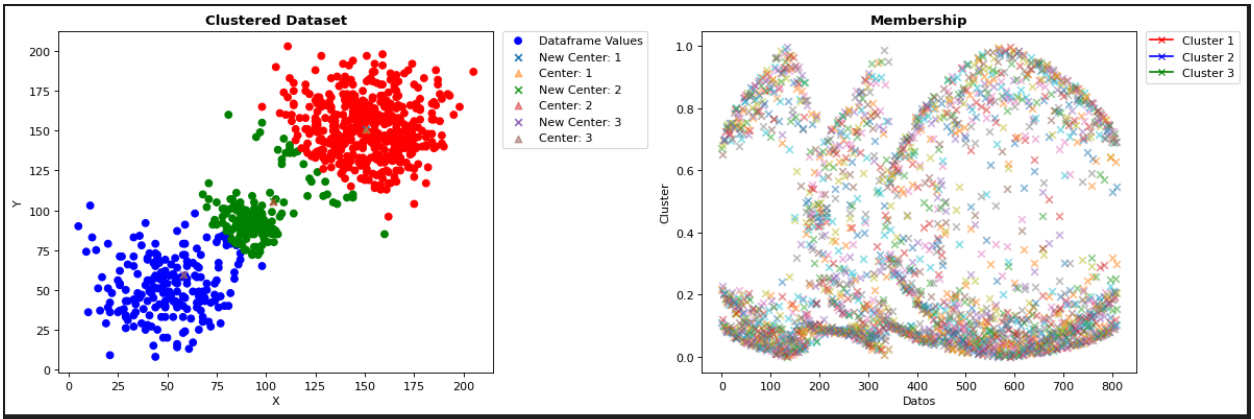

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster

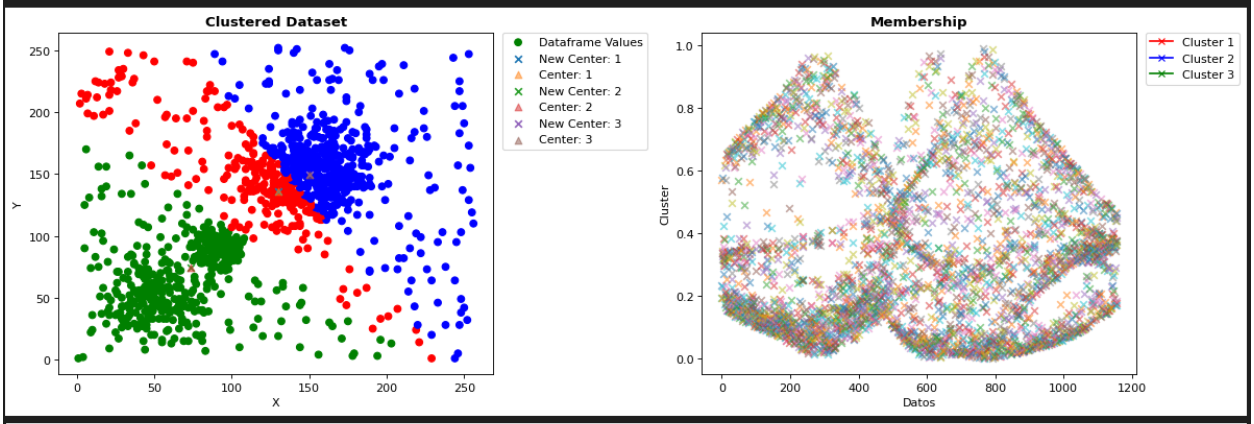

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster

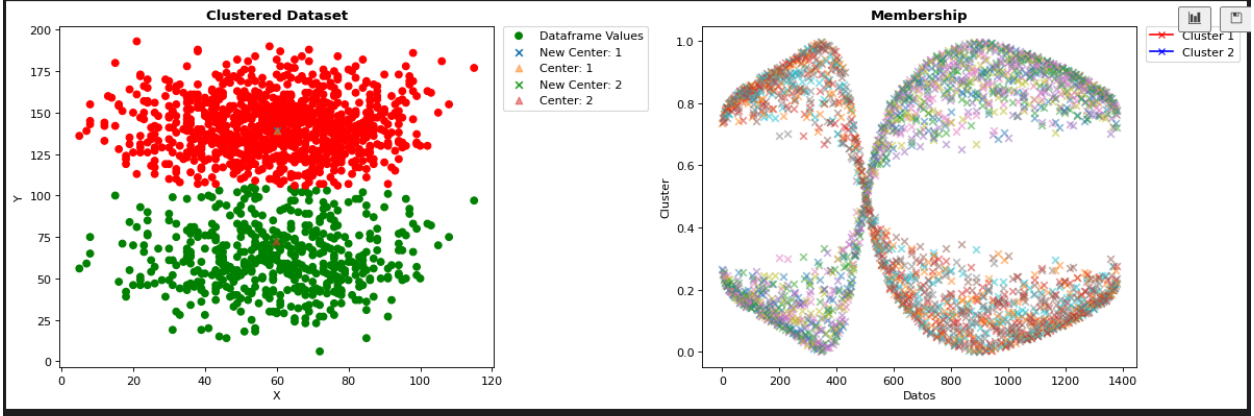

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster

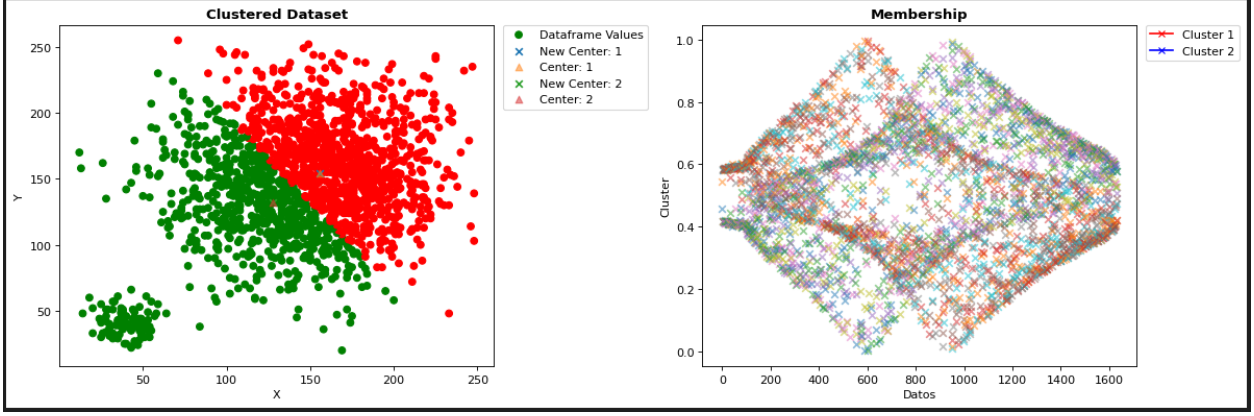
b. M = **1.5**; Initialization of Centers = **Random Elements**; Distance Method: **Euclidean**

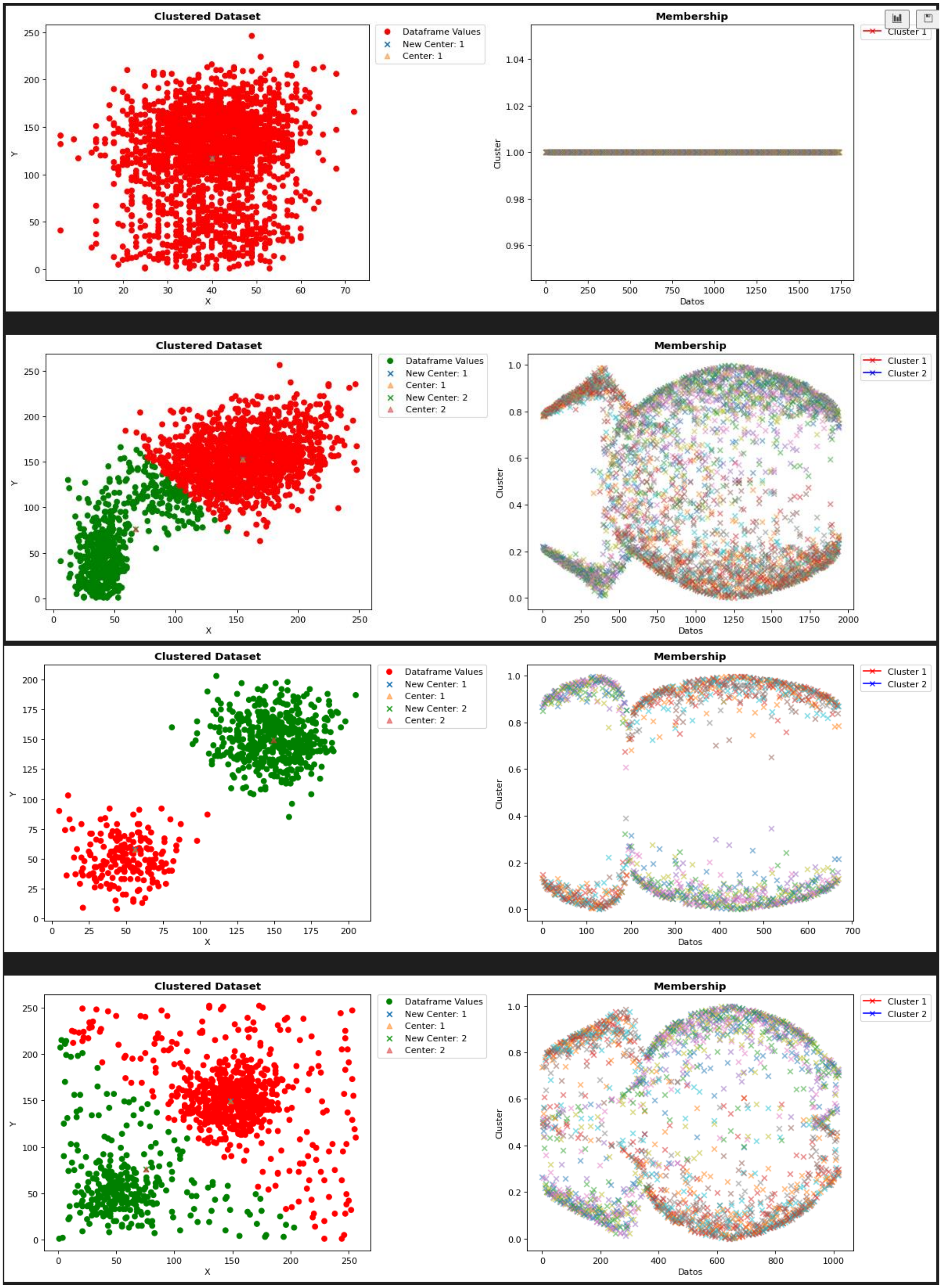



Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

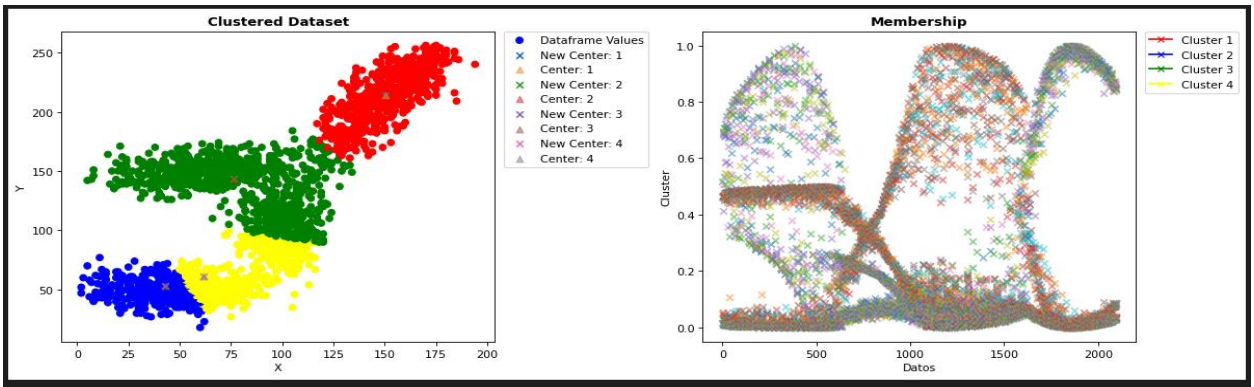

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

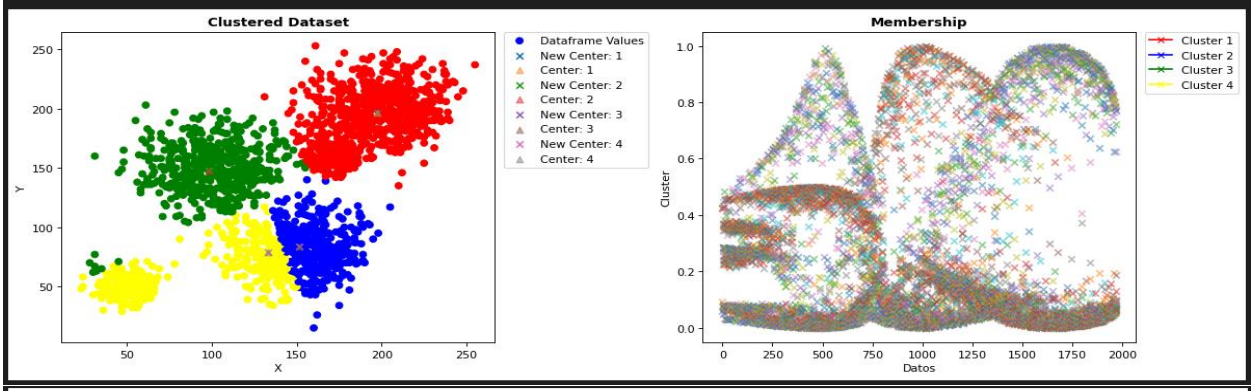

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

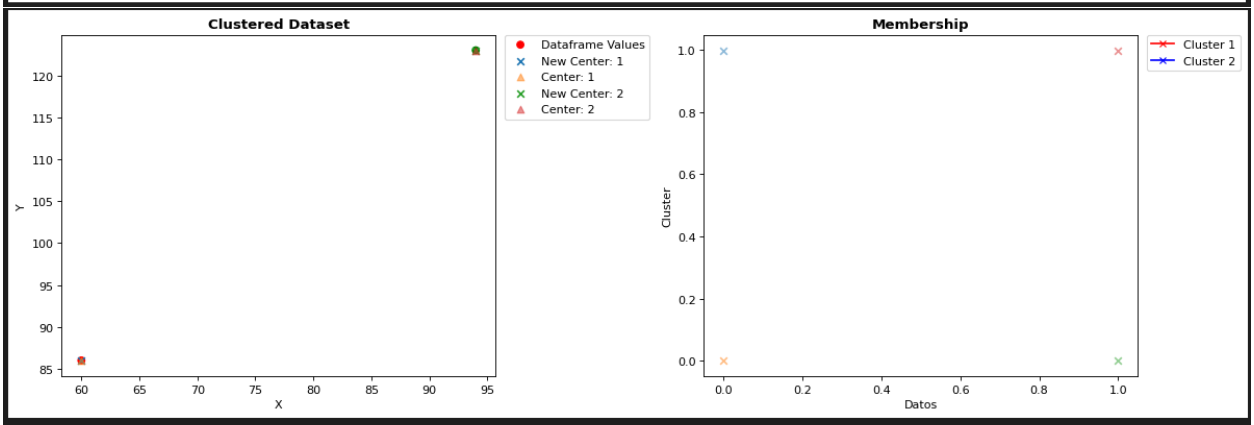

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

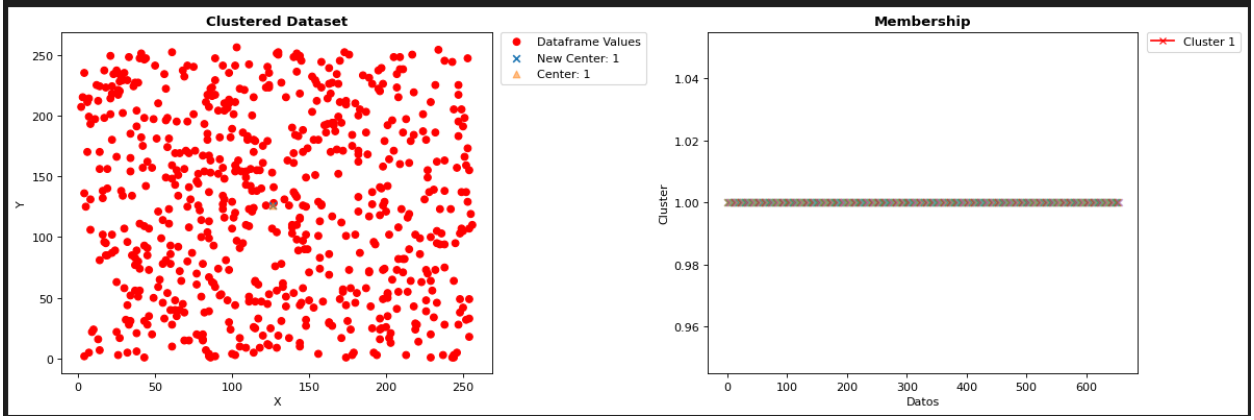

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
X
Y
Membership
Cluster 1
Datos
Cluster
Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster
Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster
Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster

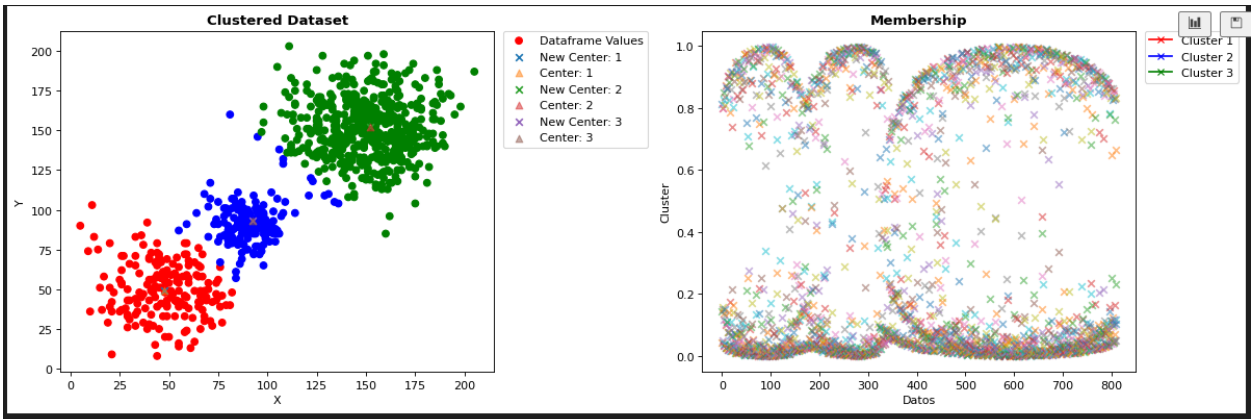
c. M = **2**; Initialization of Centers = **First Elements**; Distance Method: **Euclidean**

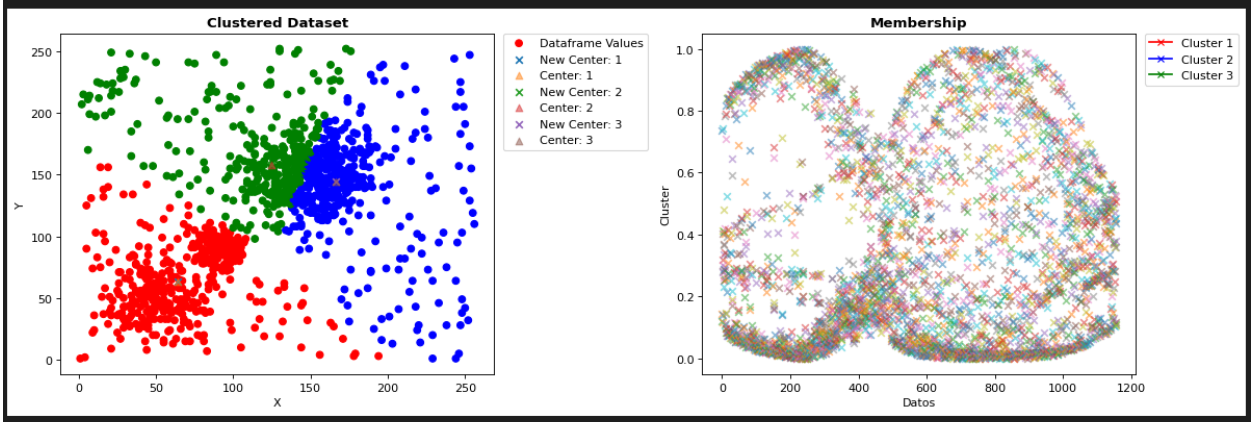


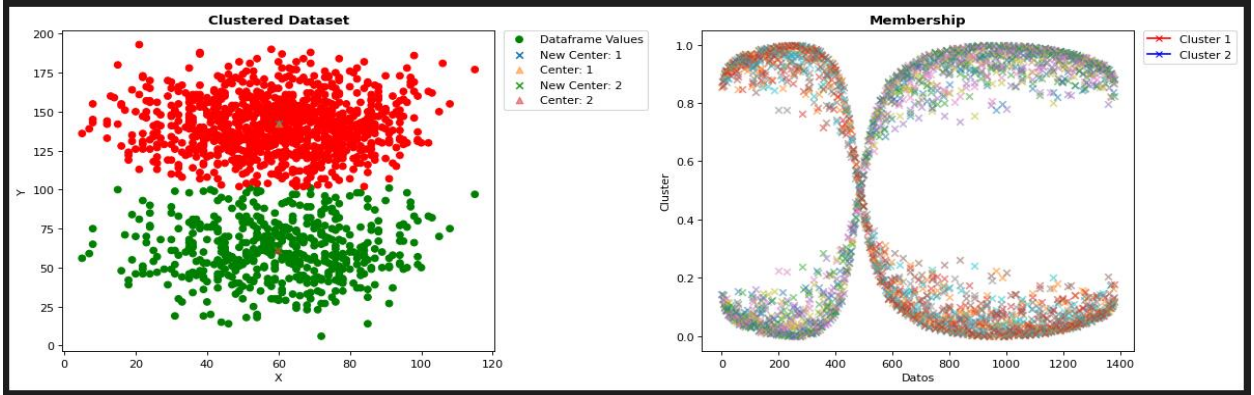


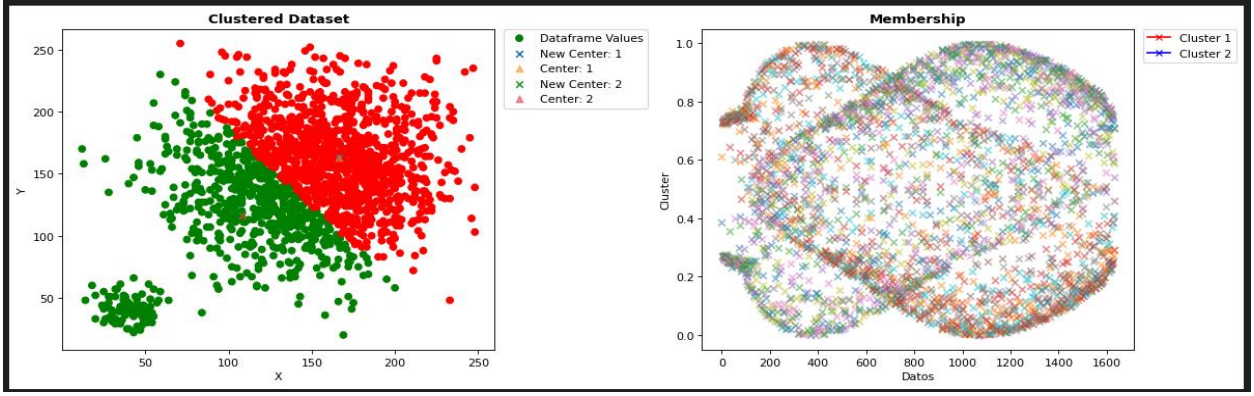


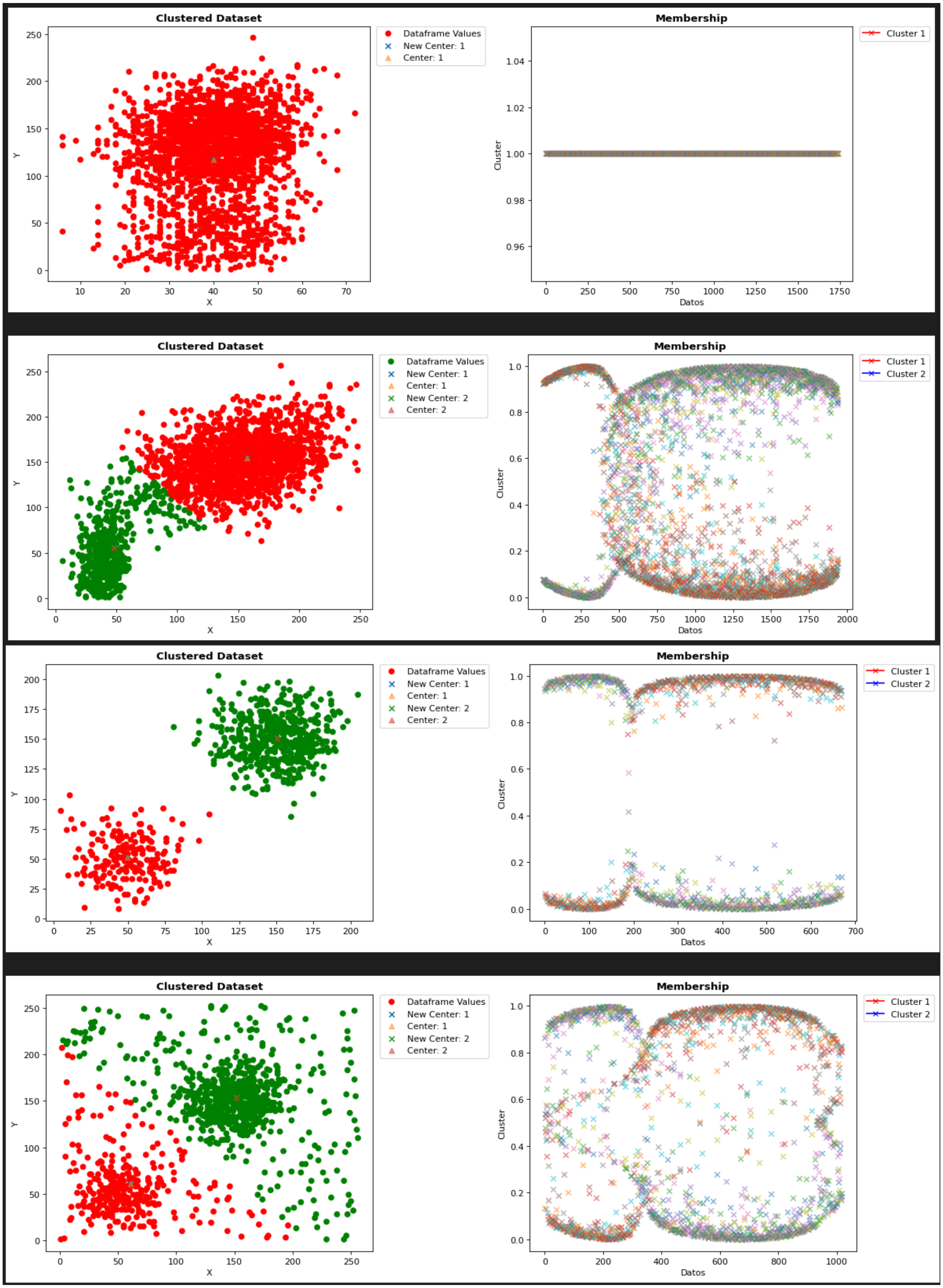



Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos


Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

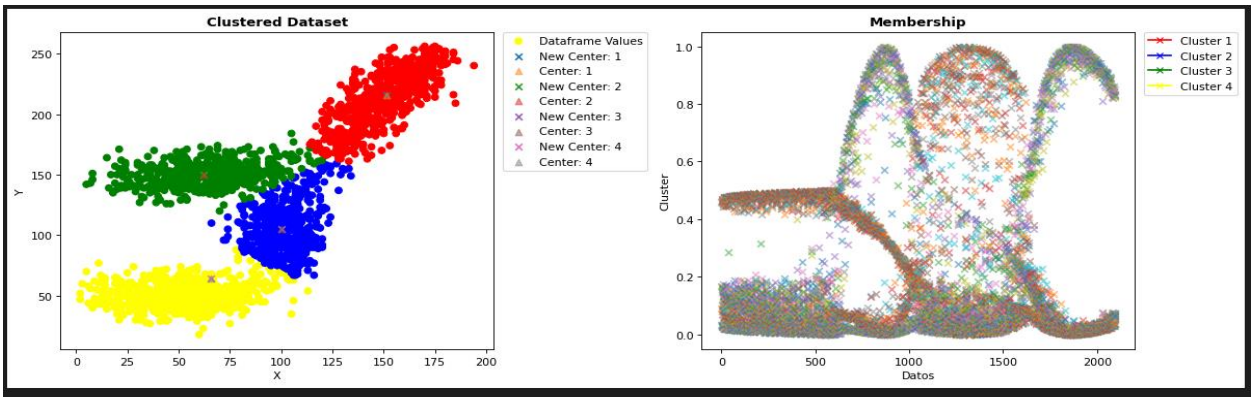

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

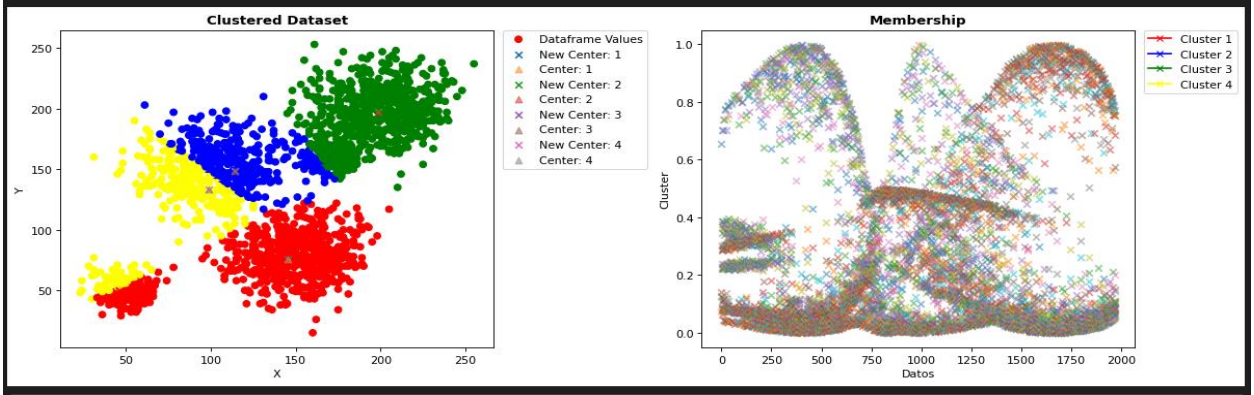

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

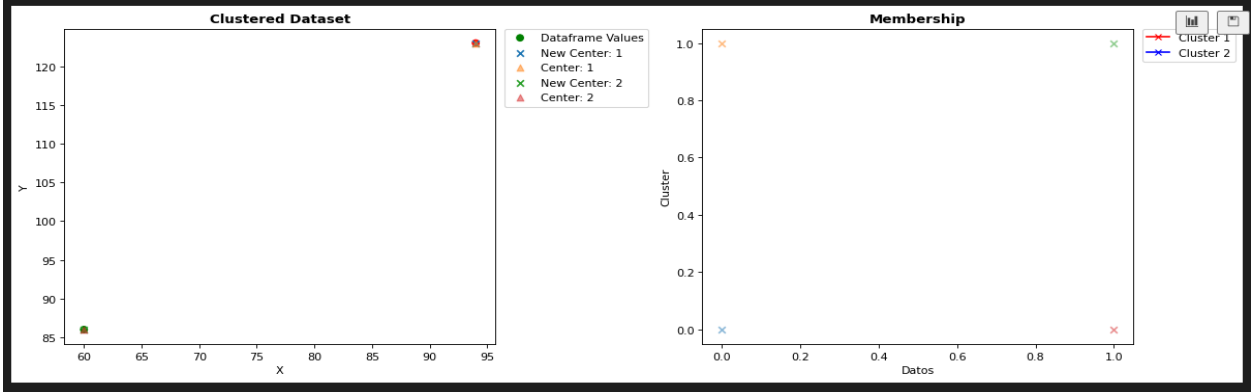


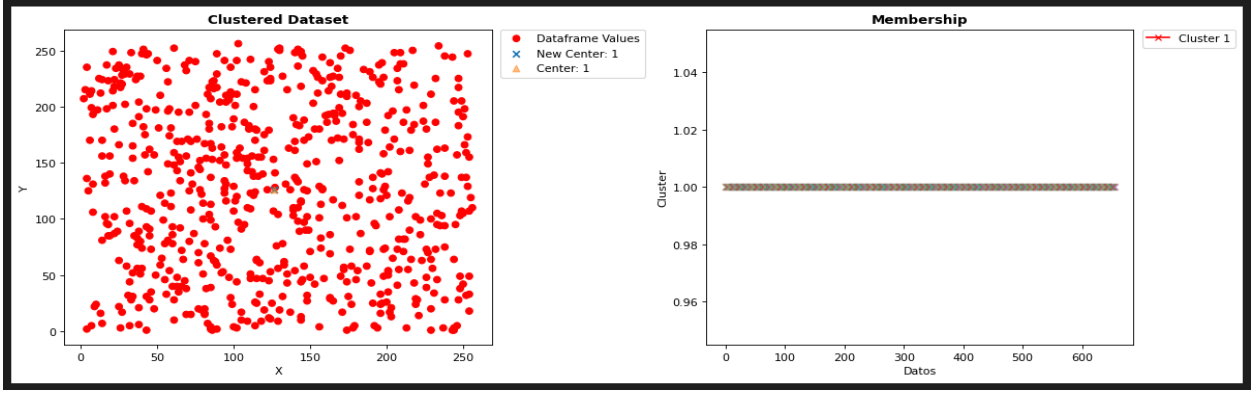
Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
X
Y
Membership
Cluster 1
Datos
Cluster
Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster
Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster
Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster

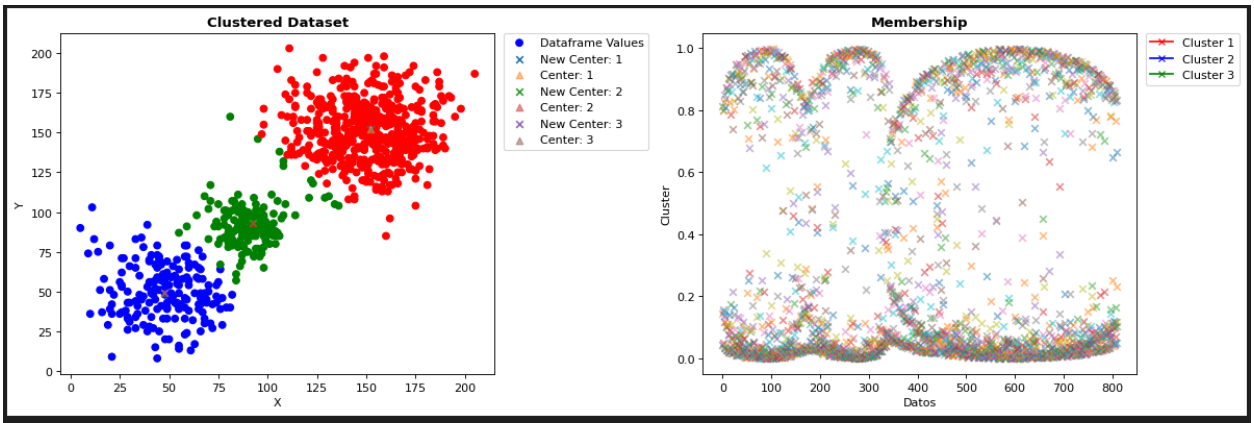
d. M = **2**; Initialization of Centers = **Random Elements**; Distance Method: **Euclidean**

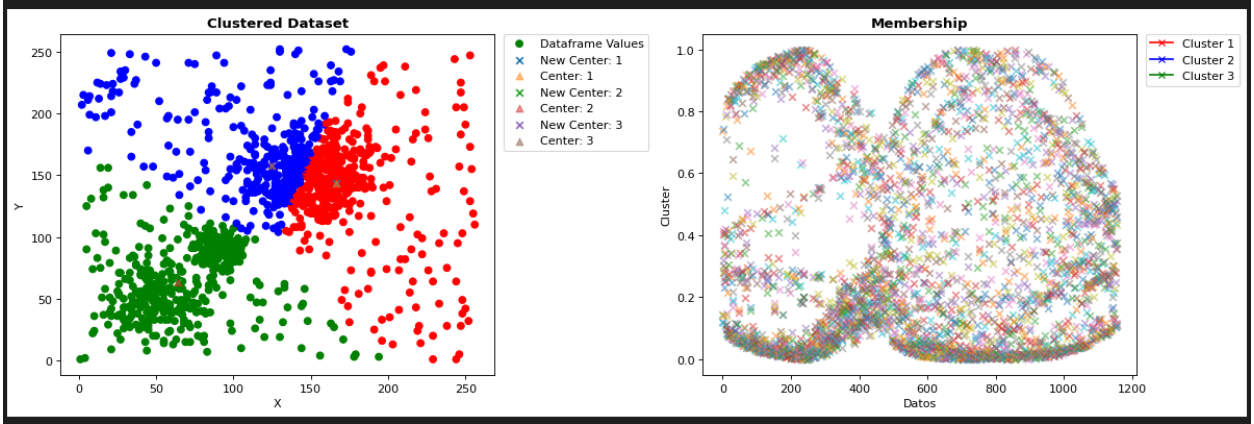


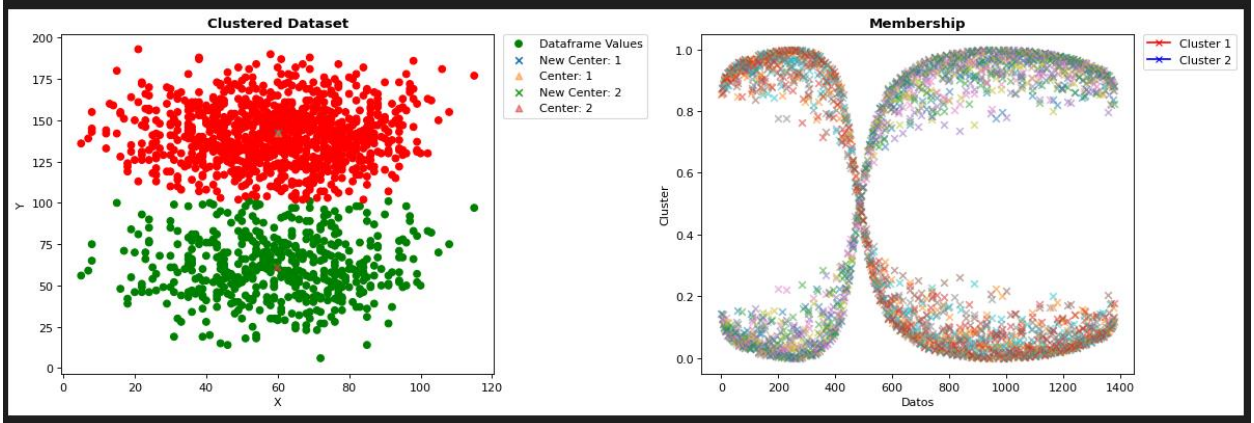


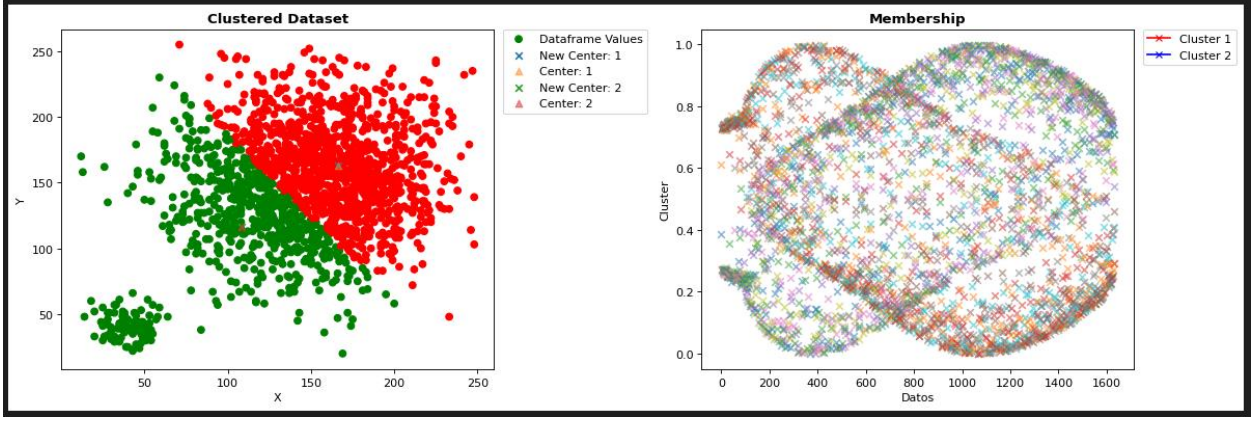


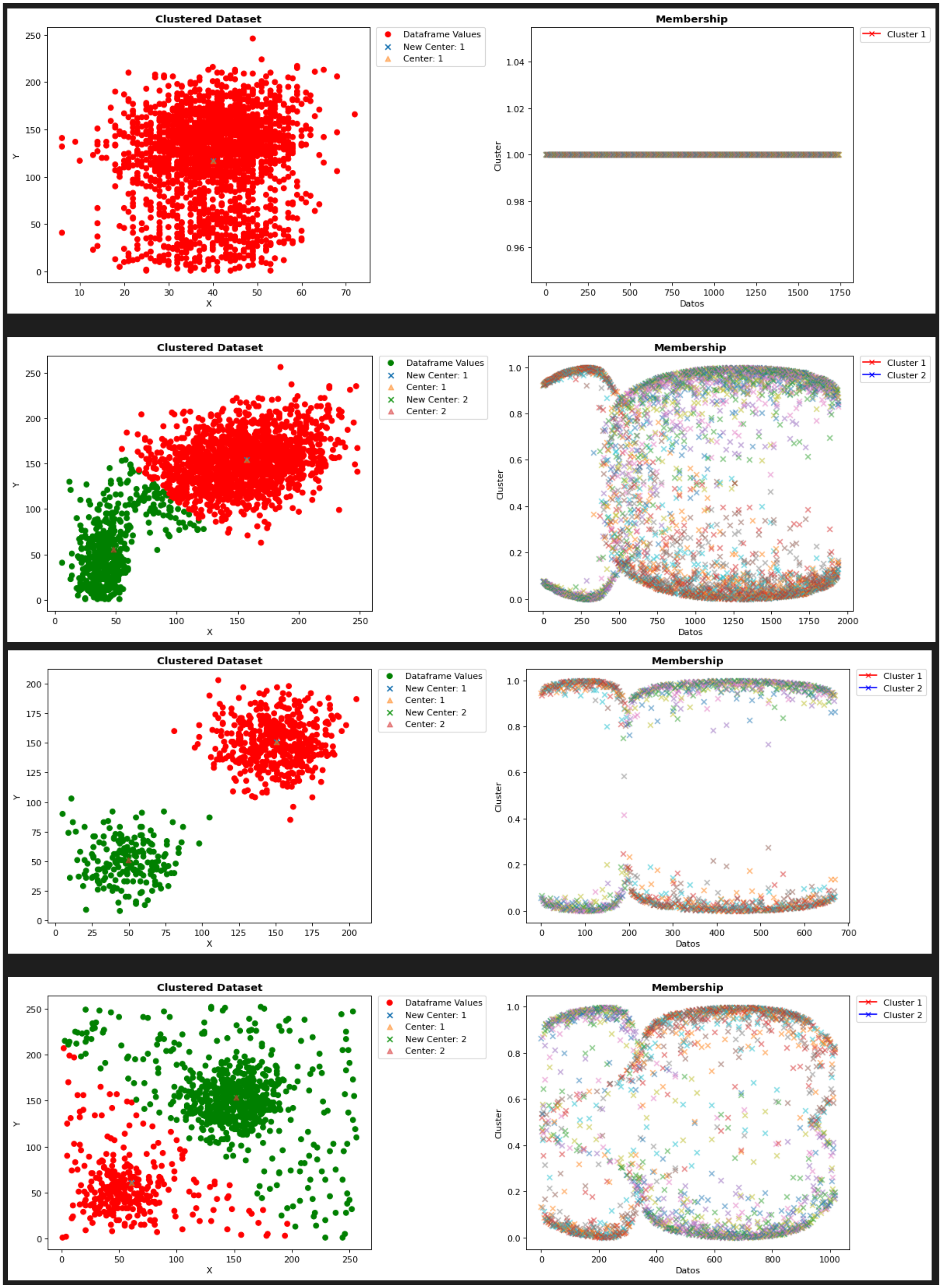



Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

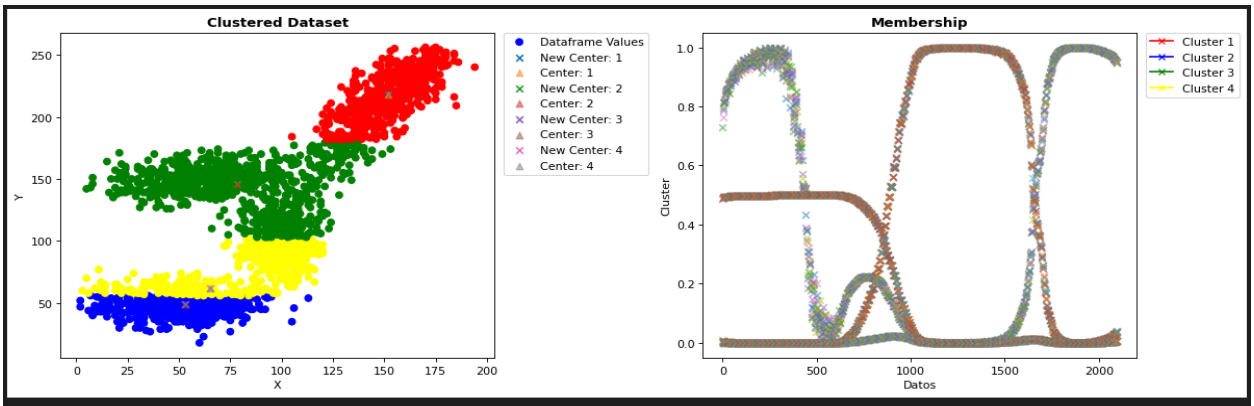

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

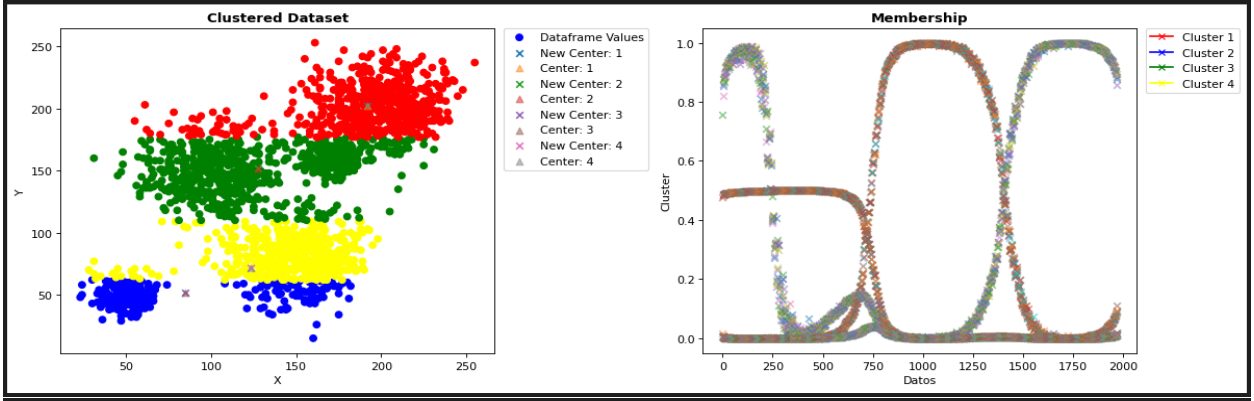

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

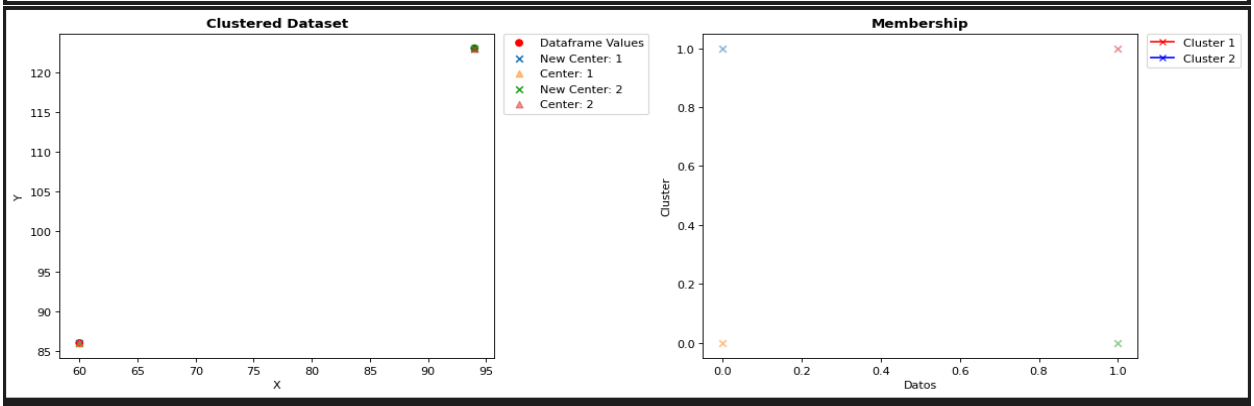

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

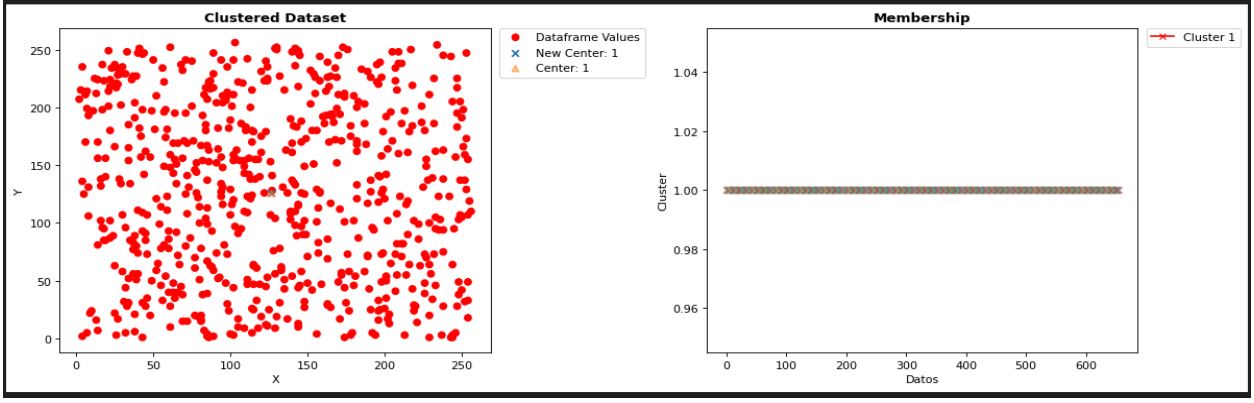

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
X
Y
Membership
Cluster 1
Datos
Cluster
Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster
Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster
Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster

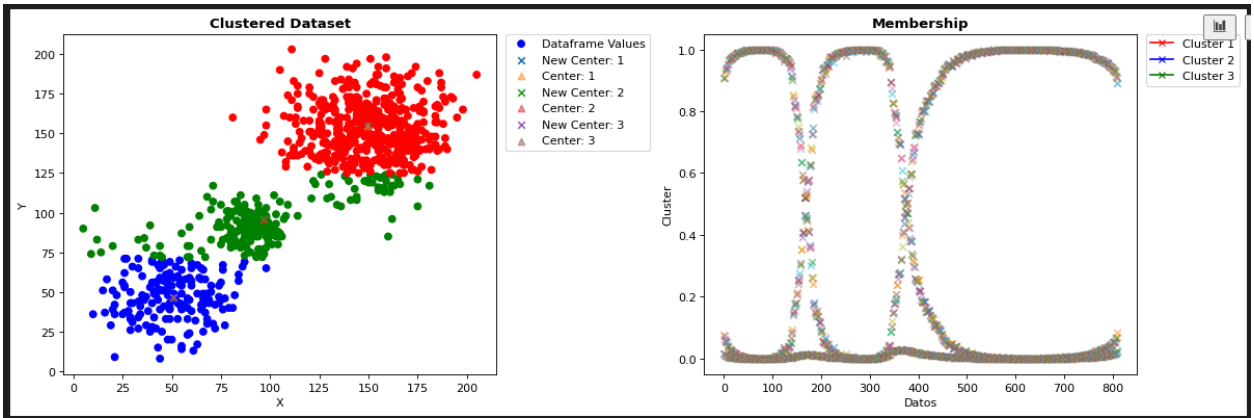
e. M = **1.5**; Initialization of Centers = **First Elements**; Distance Method: **Manhattan**

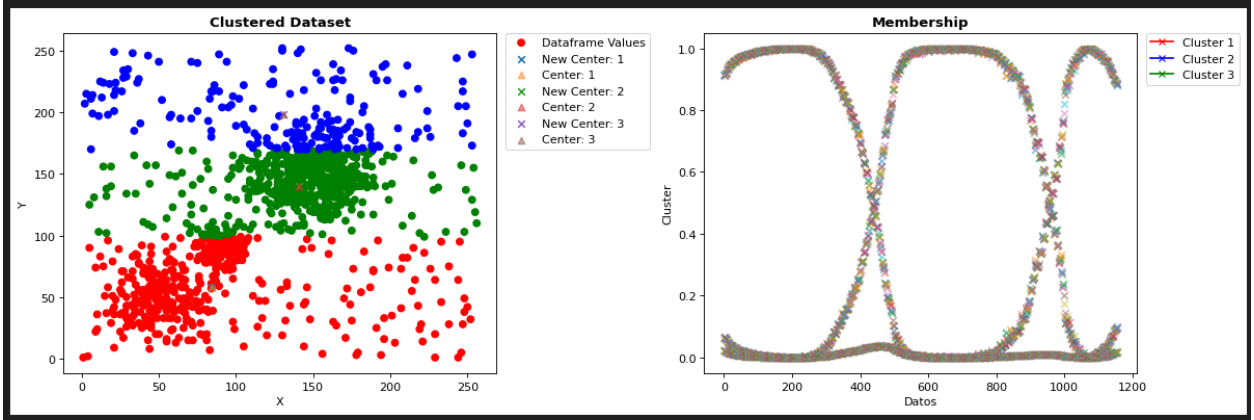


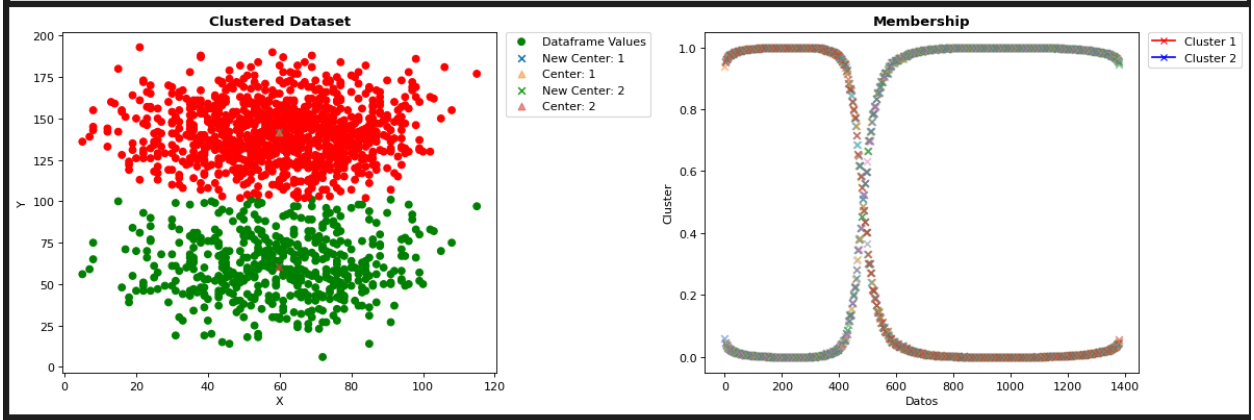


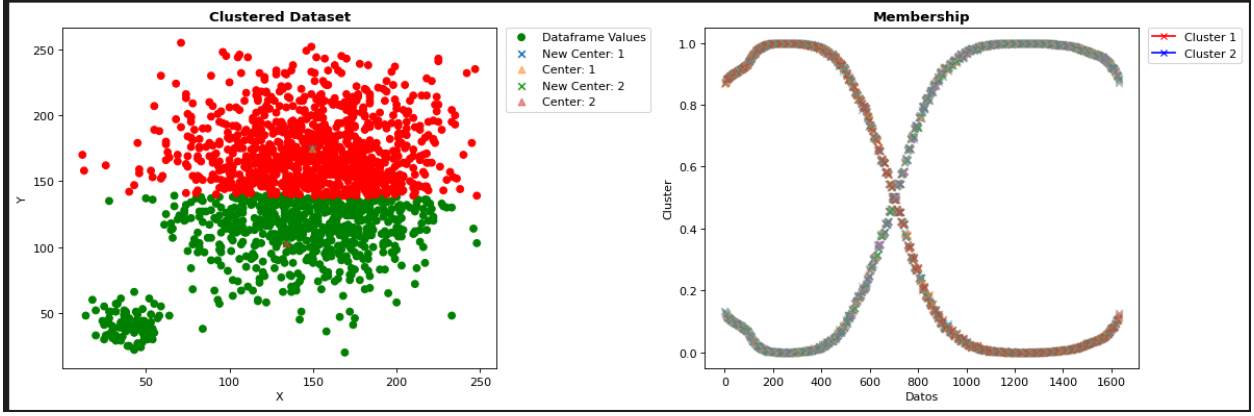


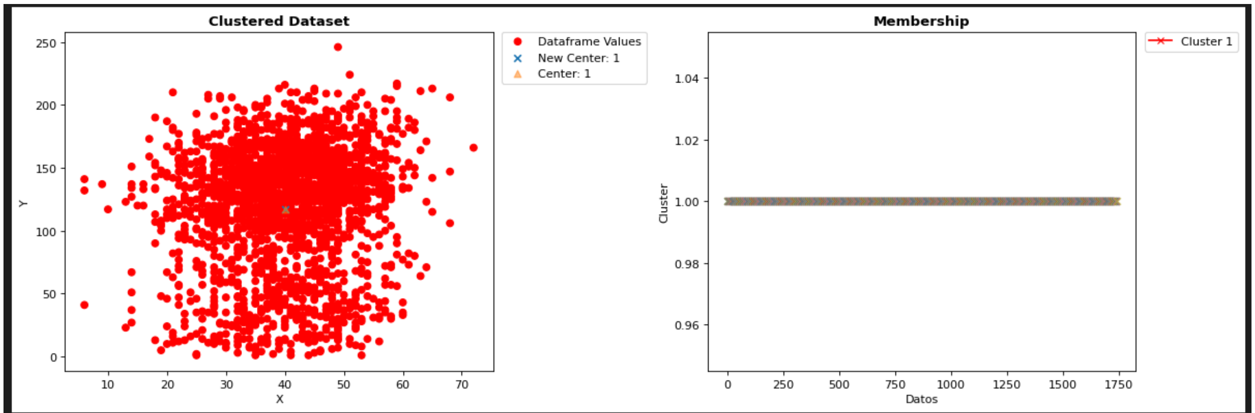


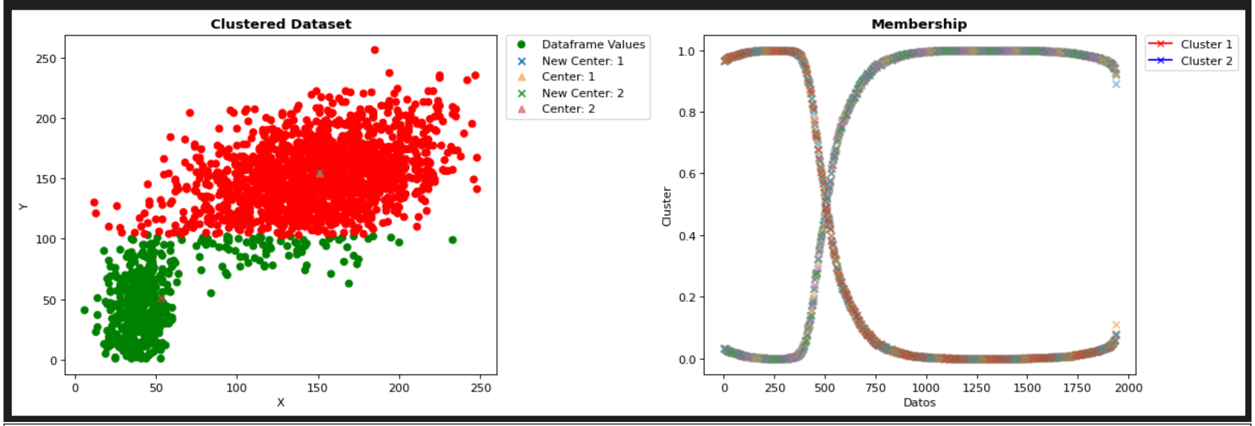

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

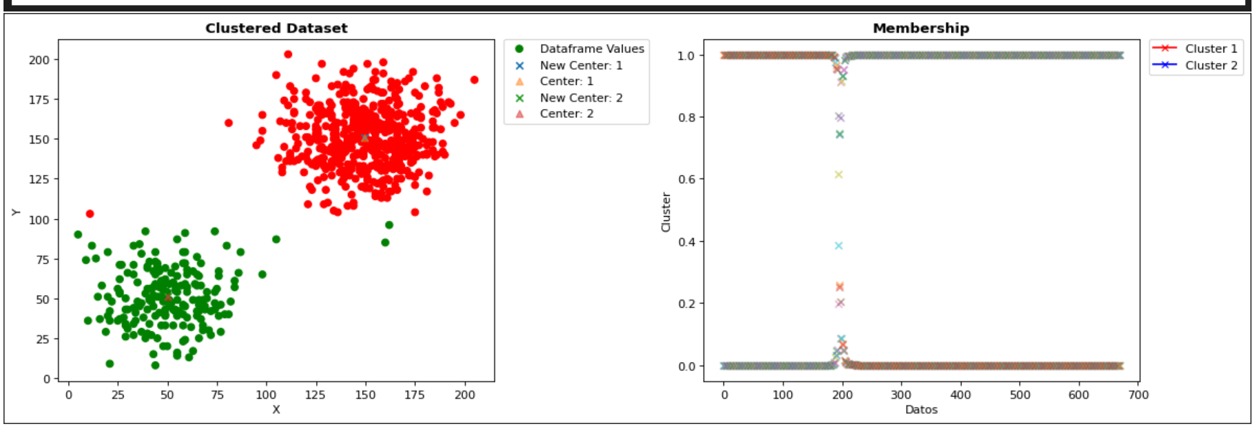

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

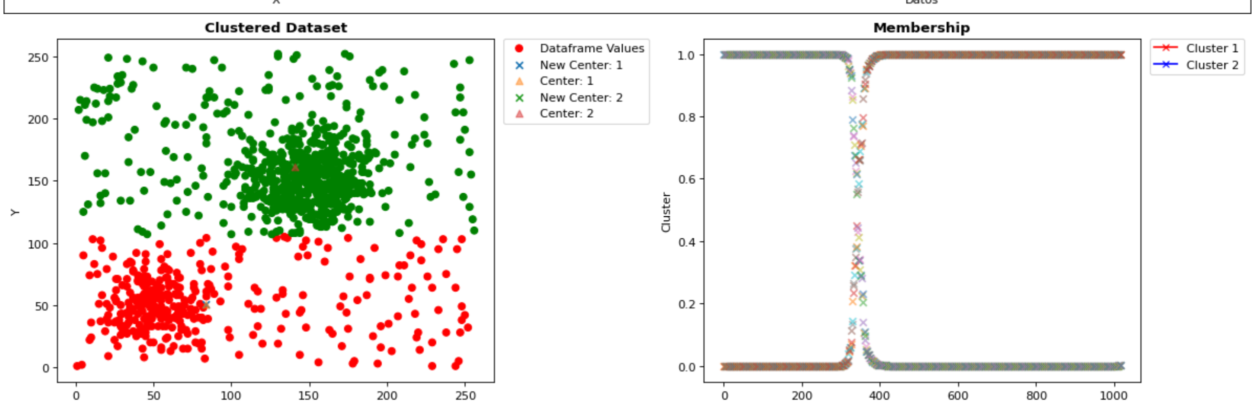

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos


Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

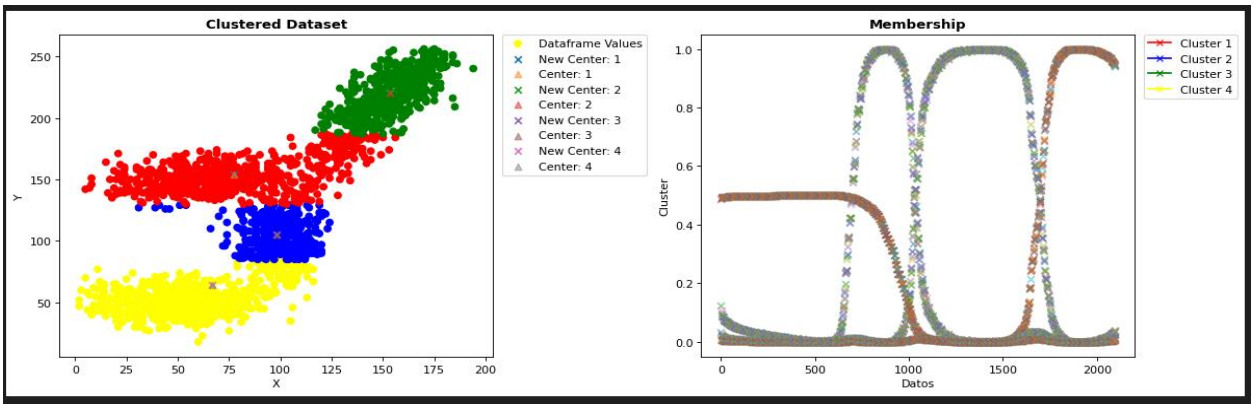

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
X
Y
Membership
Cluster 1
Datos
Cluster

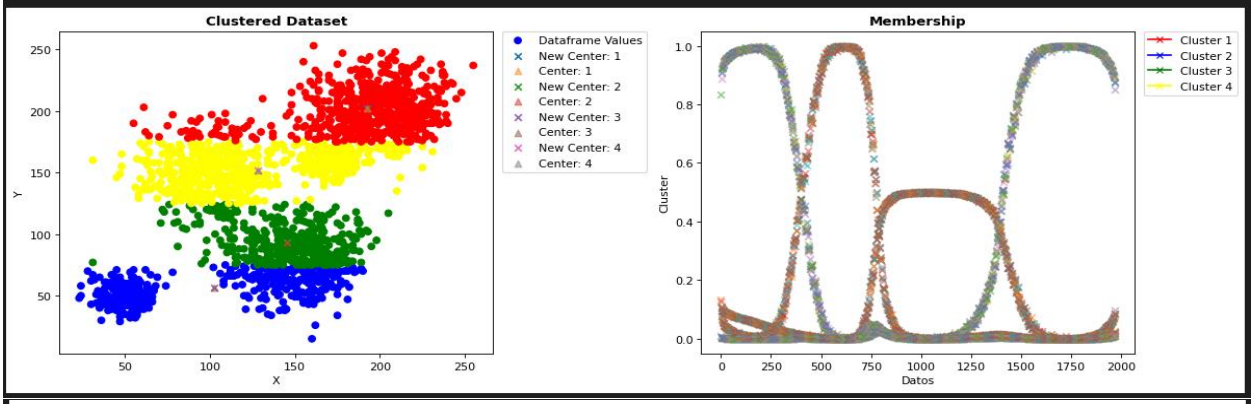

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster

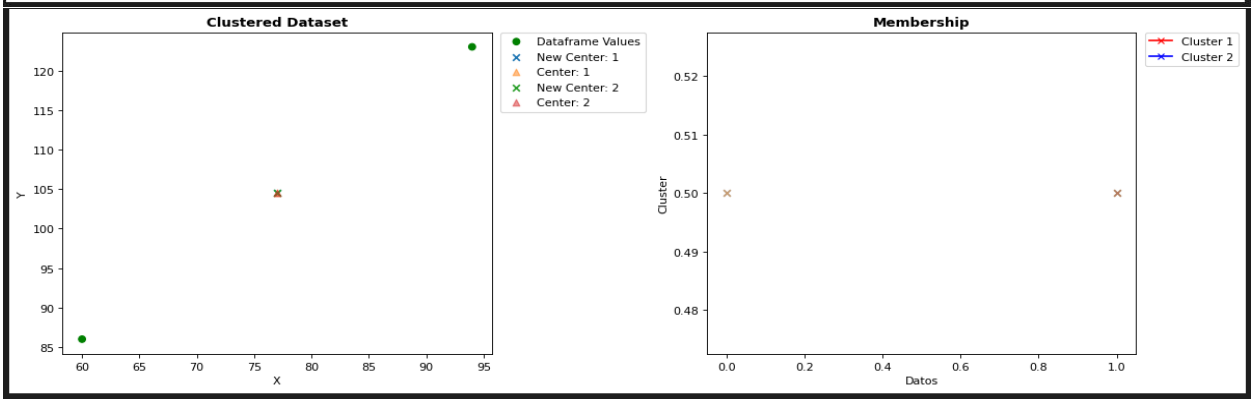

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster

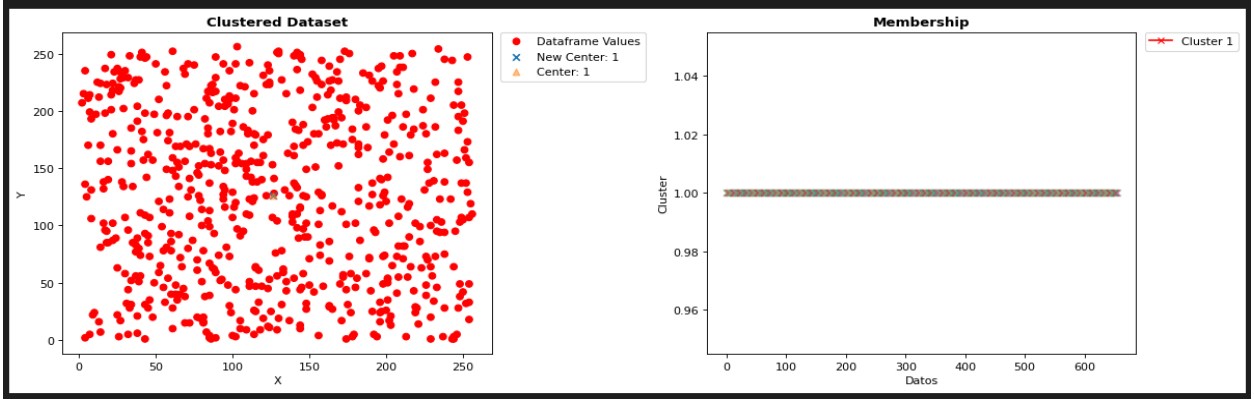

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster

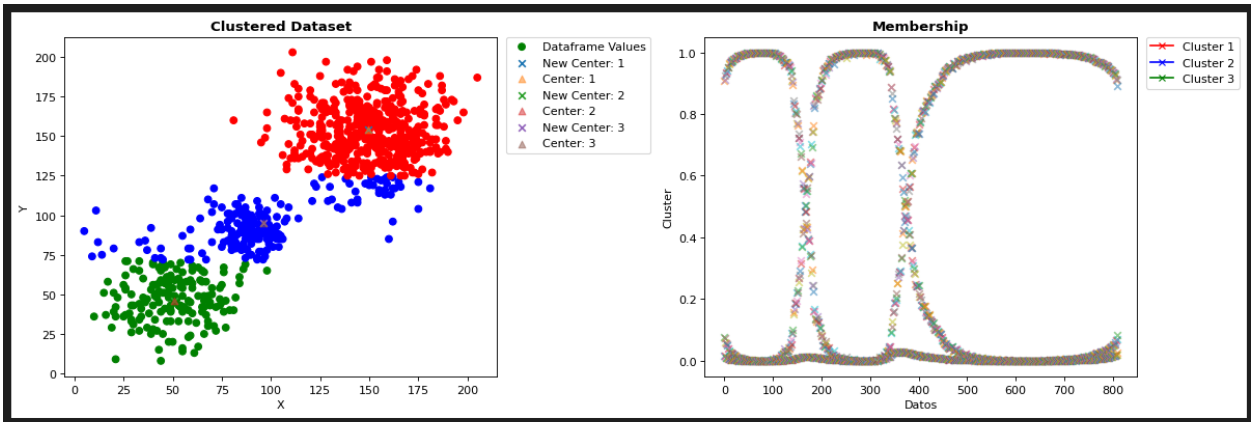
f. M = **1.5**; Initialization of Centers = **Random Elements**; Distance Method: **Manhattan**

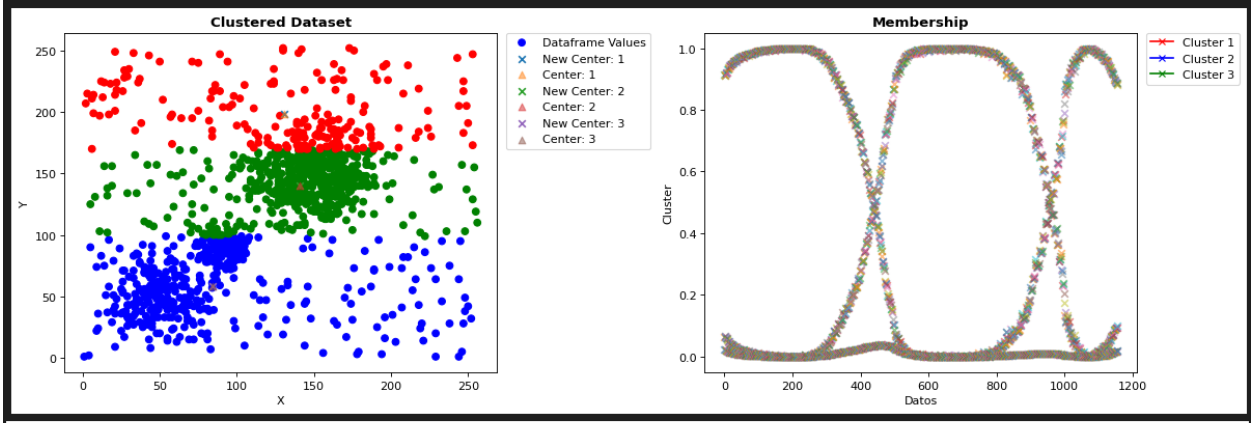


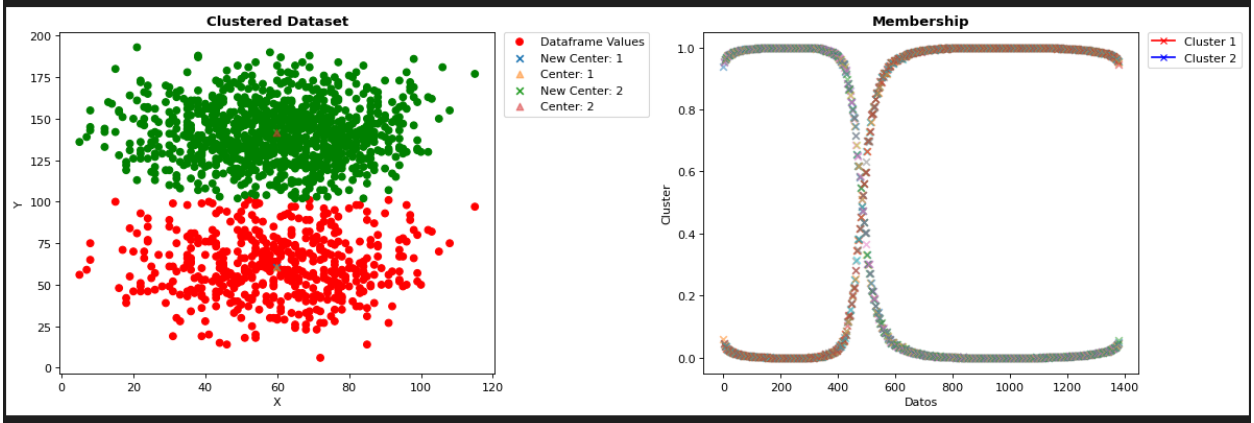


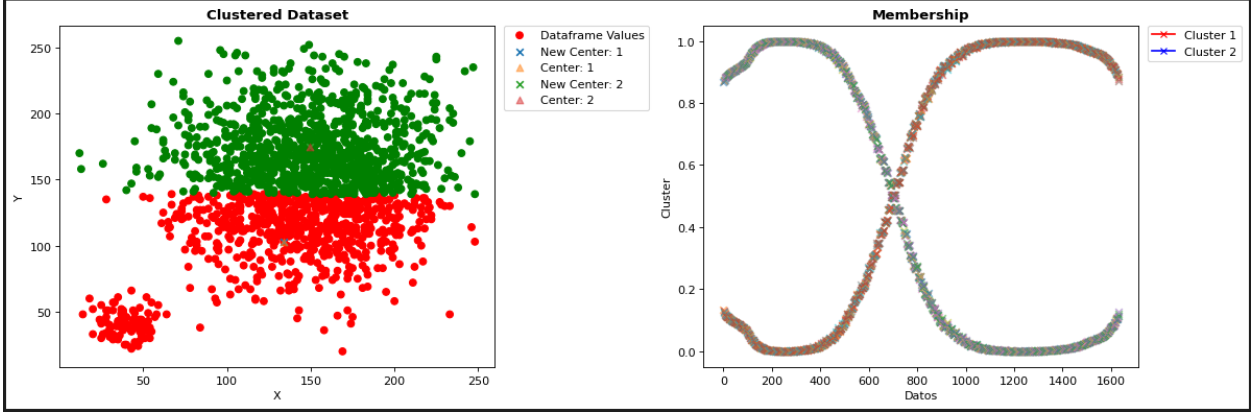


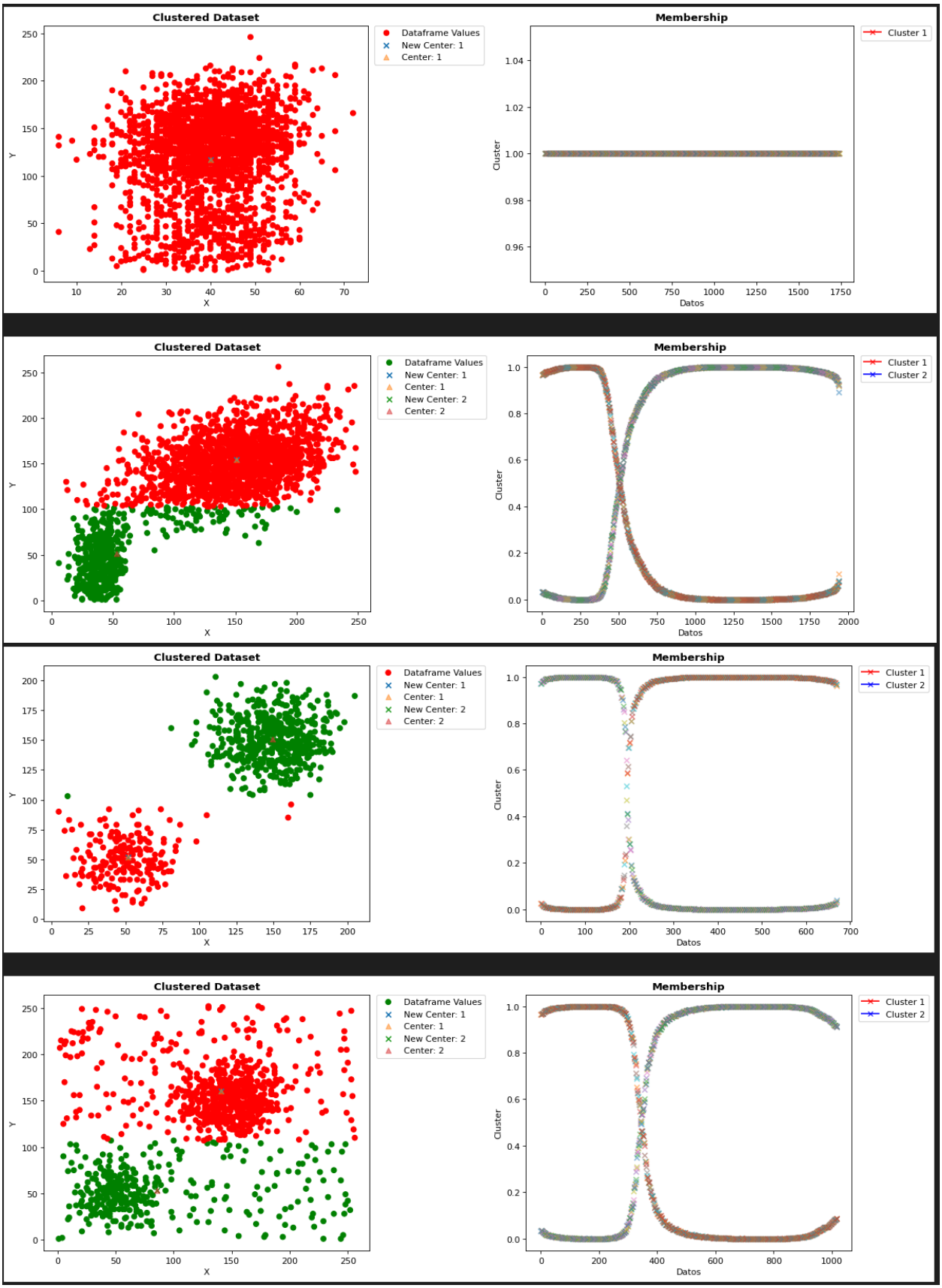



Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

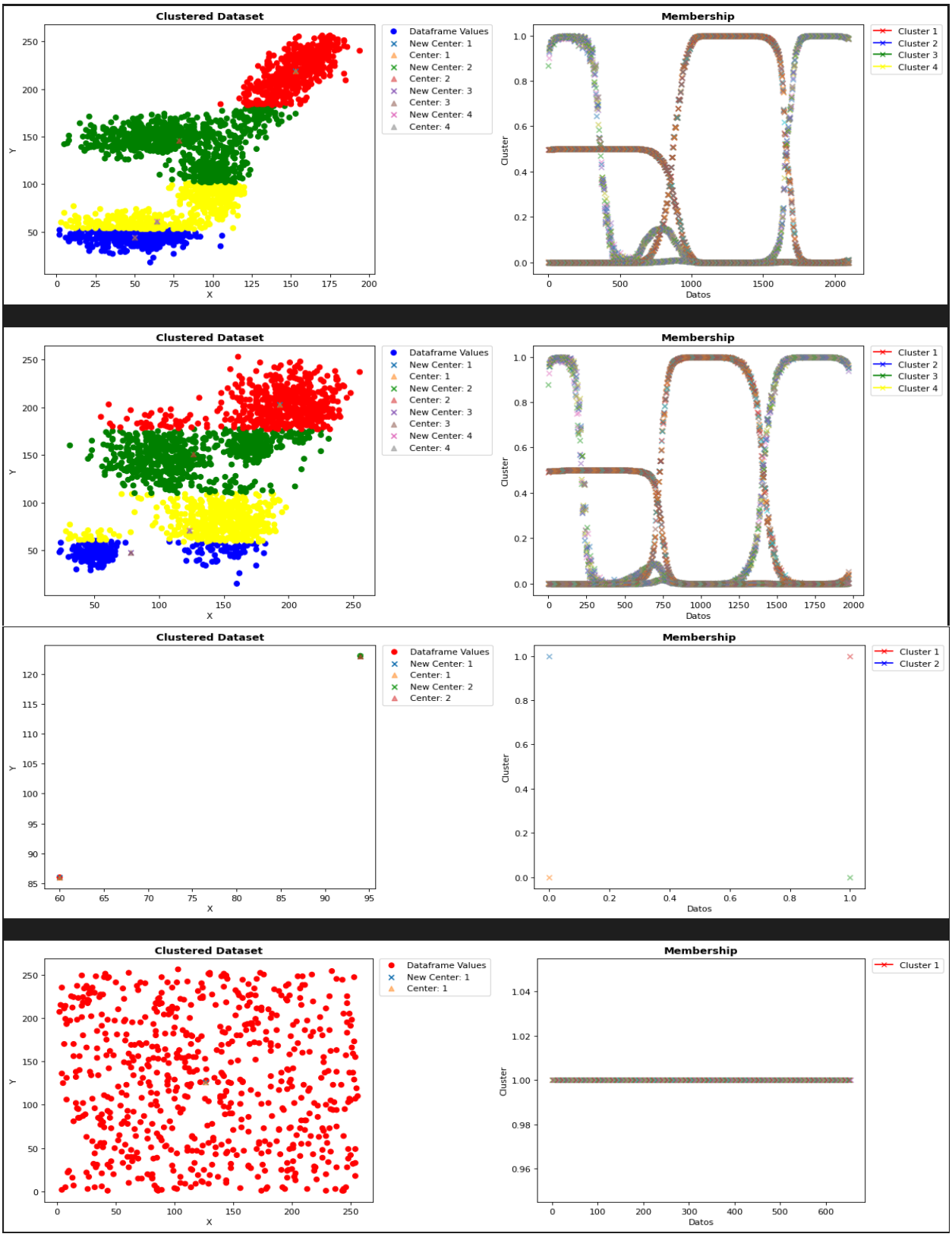

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

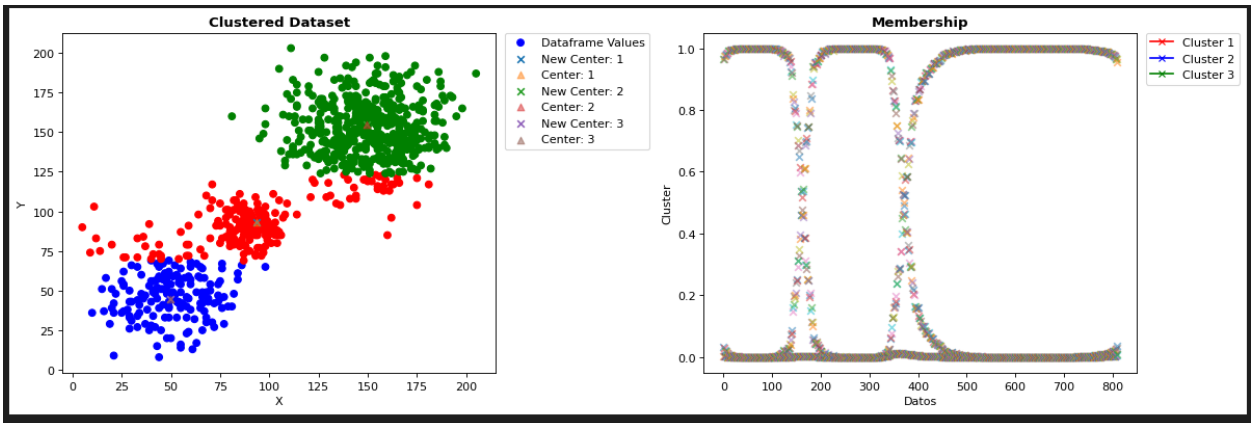

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

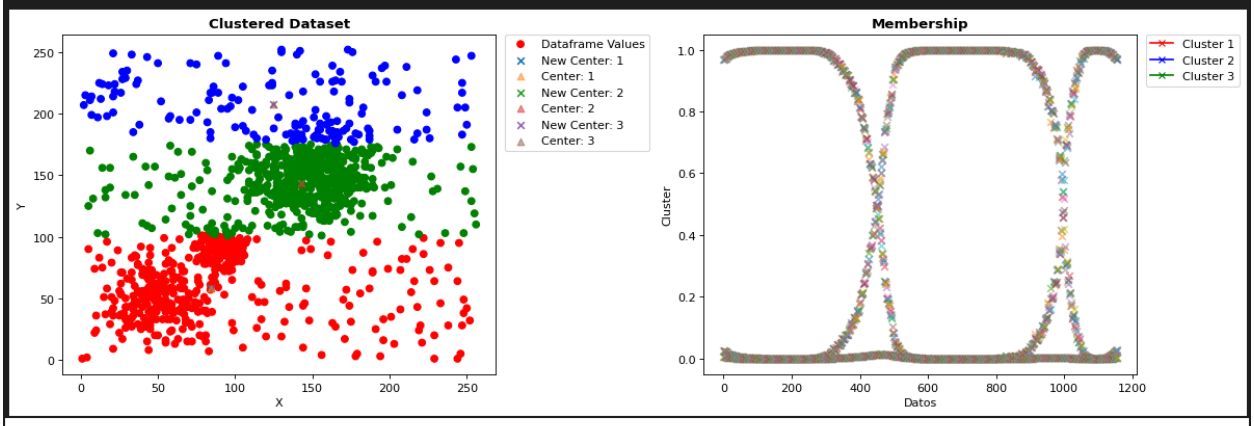

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

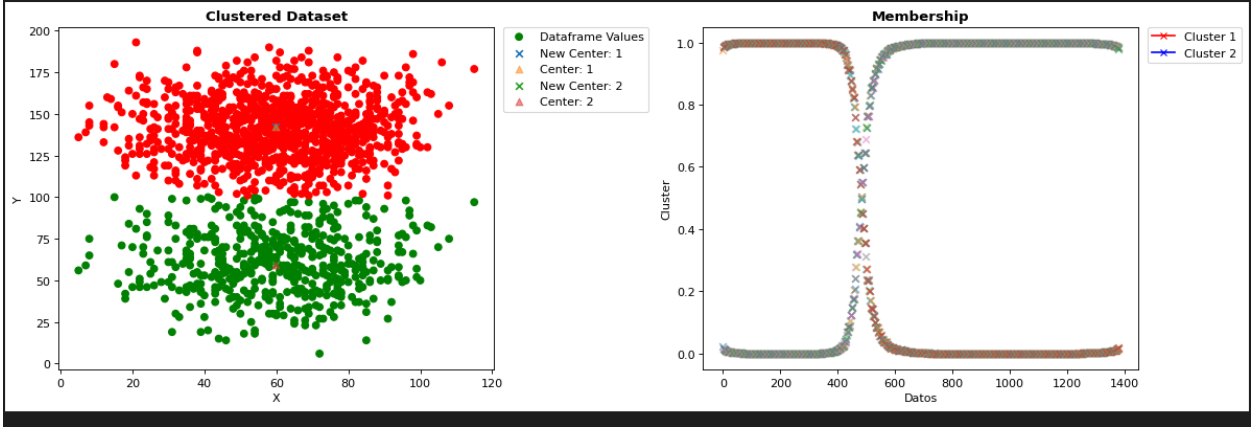

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
X
Y
Membership
Cluster 1
Cluster
Datos
Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos
Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos
Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

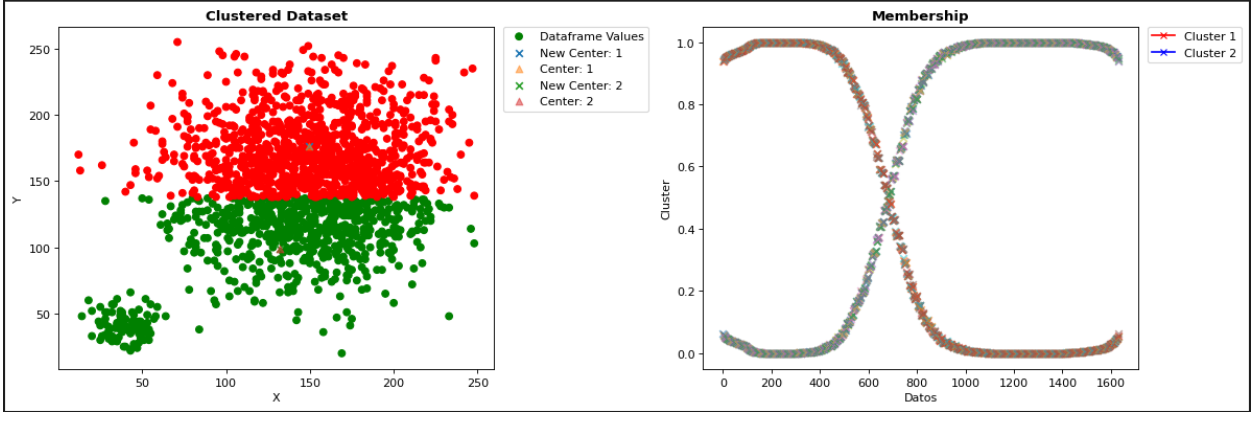
g. M = **2**; Initialization of Centers = **First Elements**; Distance Method: **Manhattan**

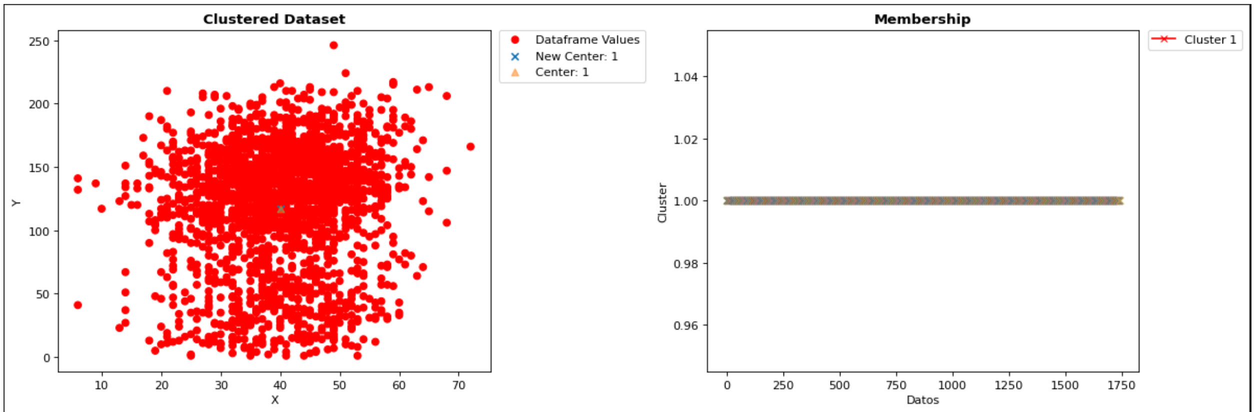


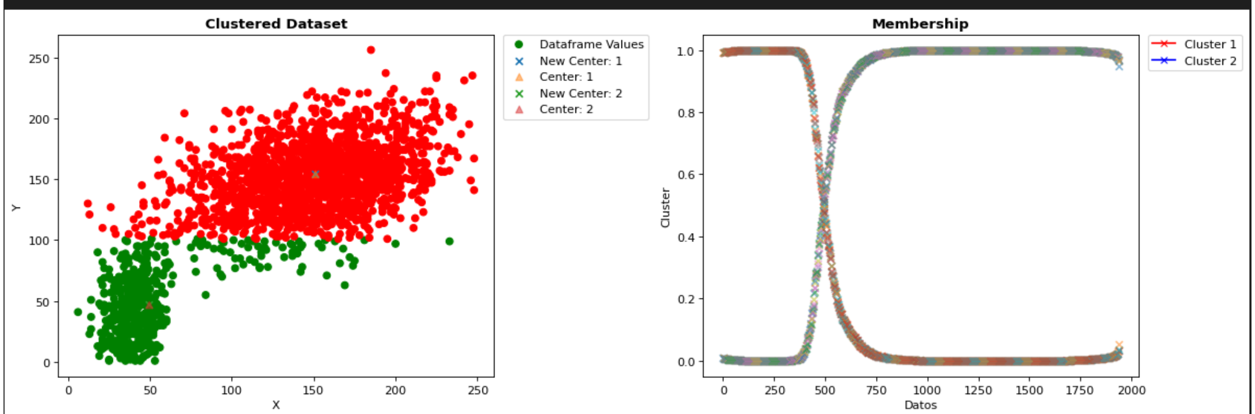

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

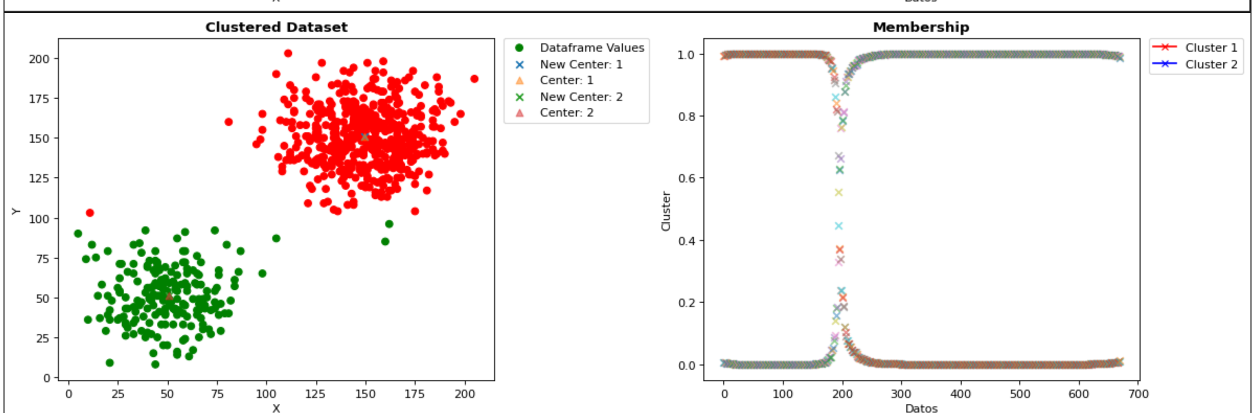

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

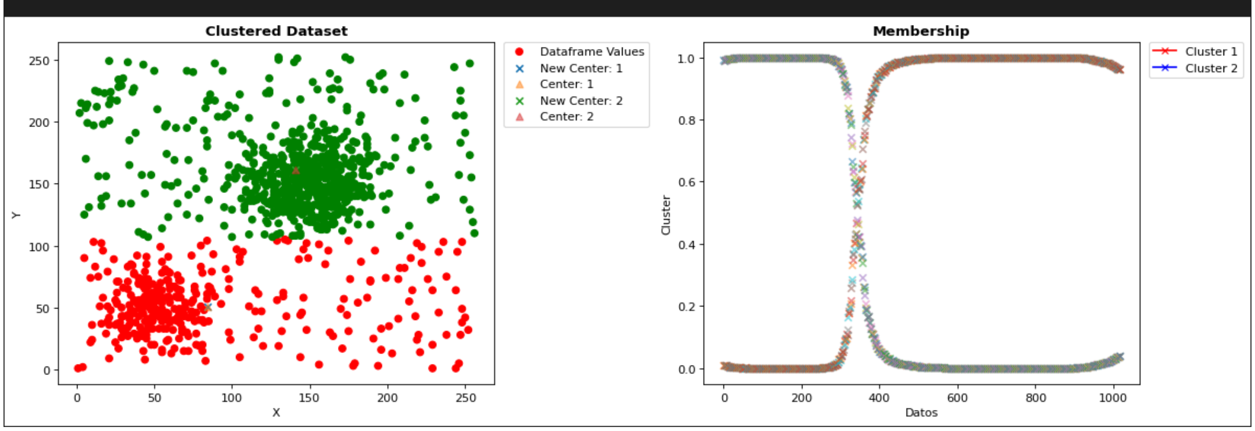

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos


Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

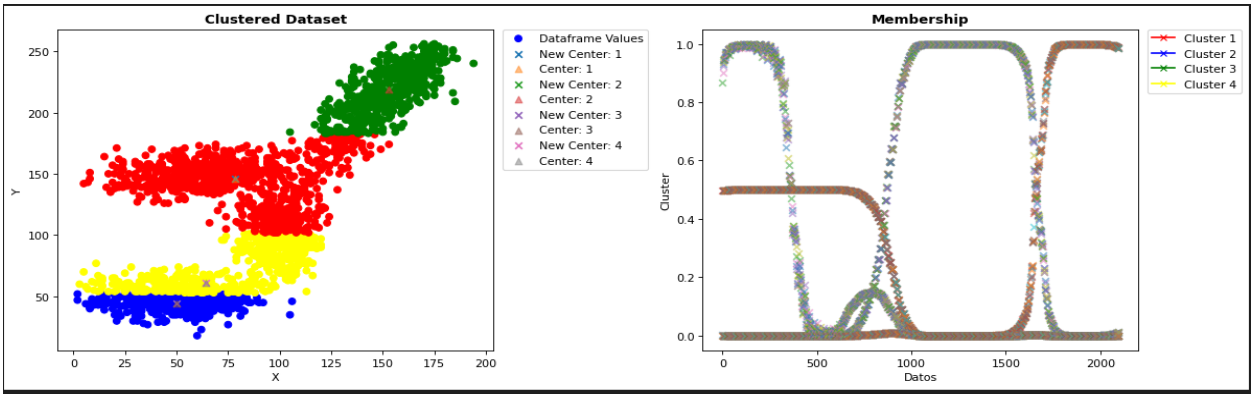

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
X
Y
Membership
Cluster 1
Cluster
Datos

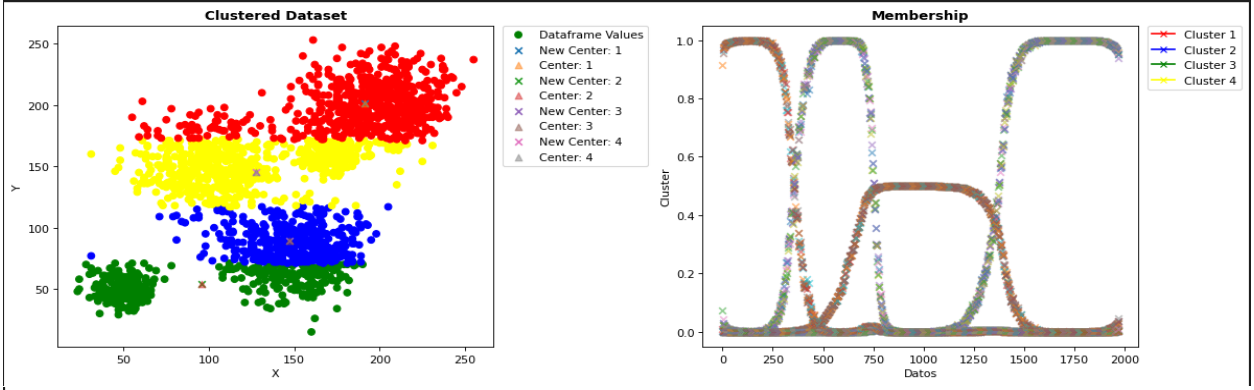

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

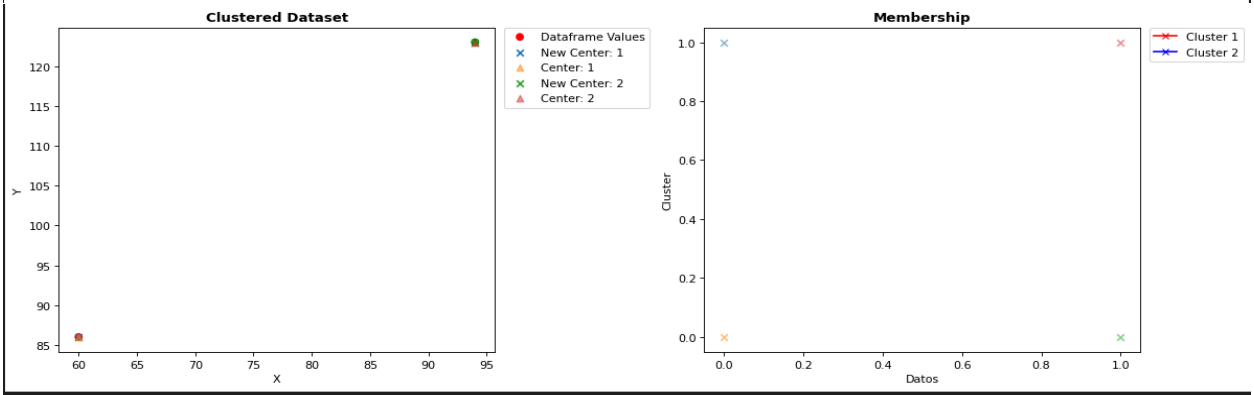

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

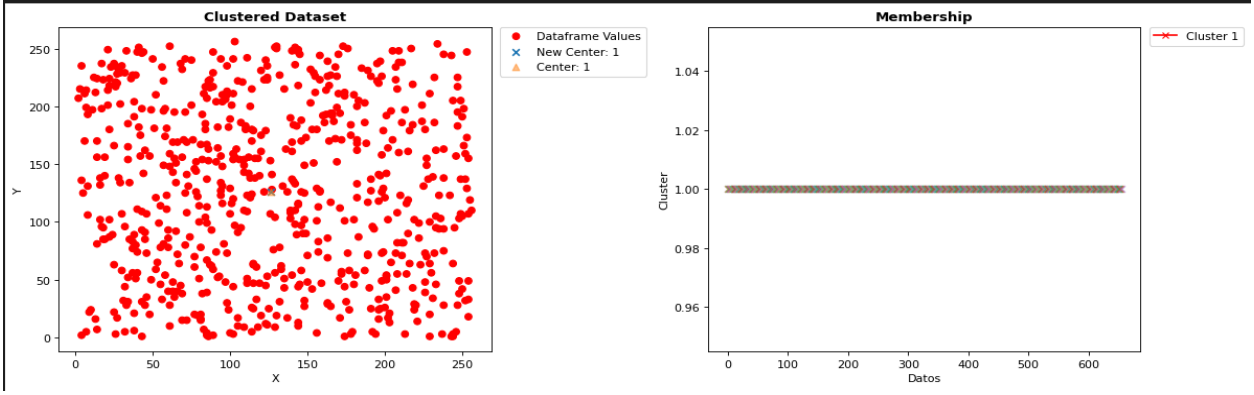

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

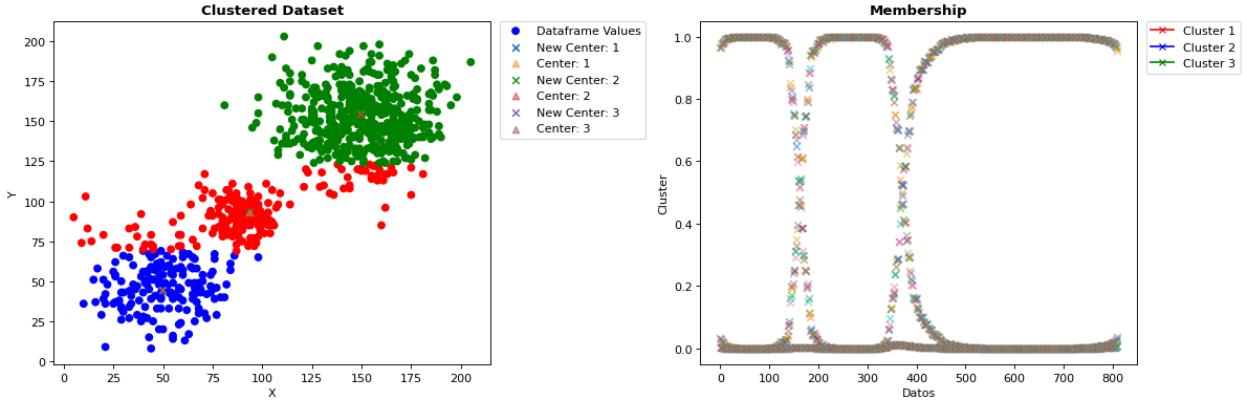
h. M = **2**; Initialization of Centers = **Random Elements**; Distance Method: **Manhattan**

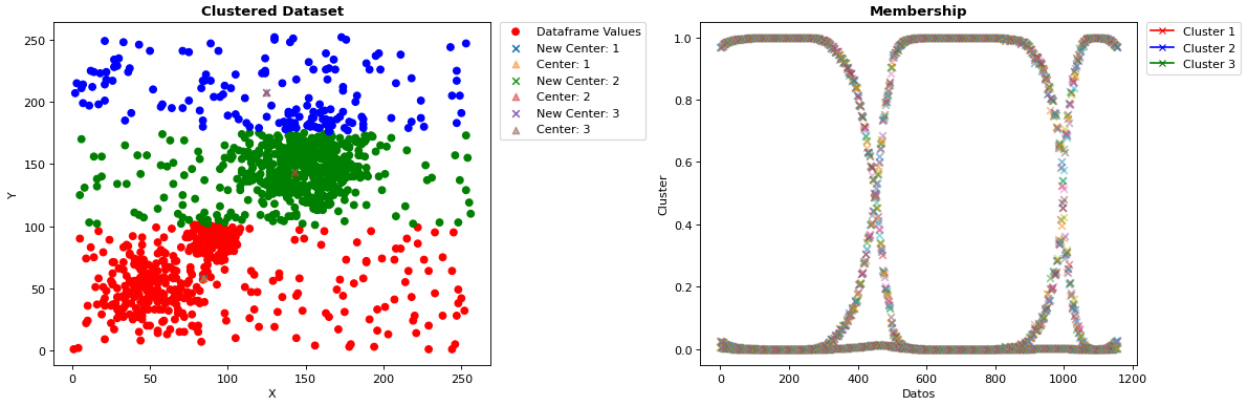


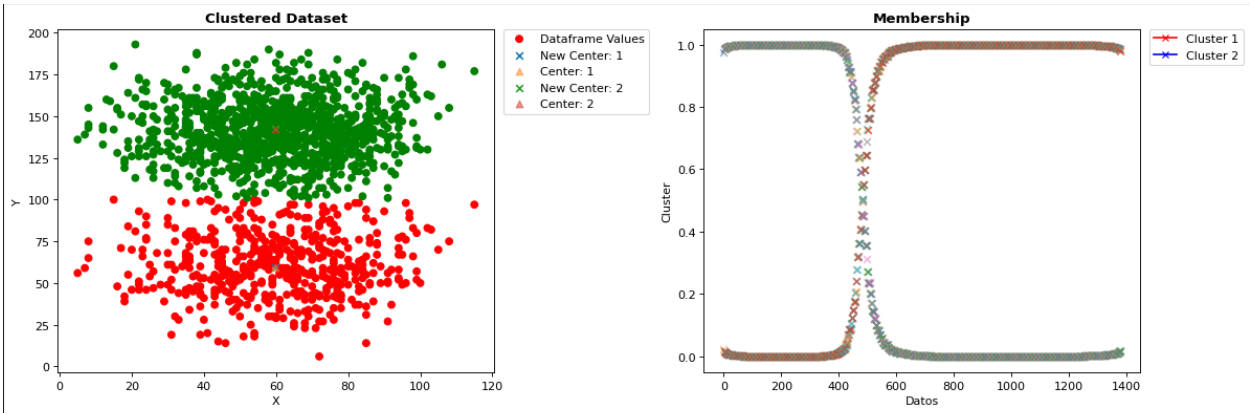


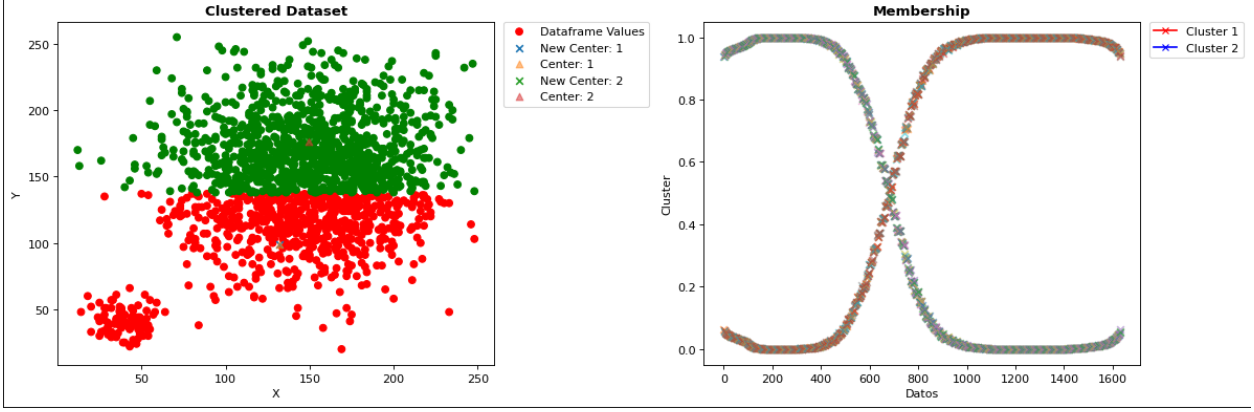


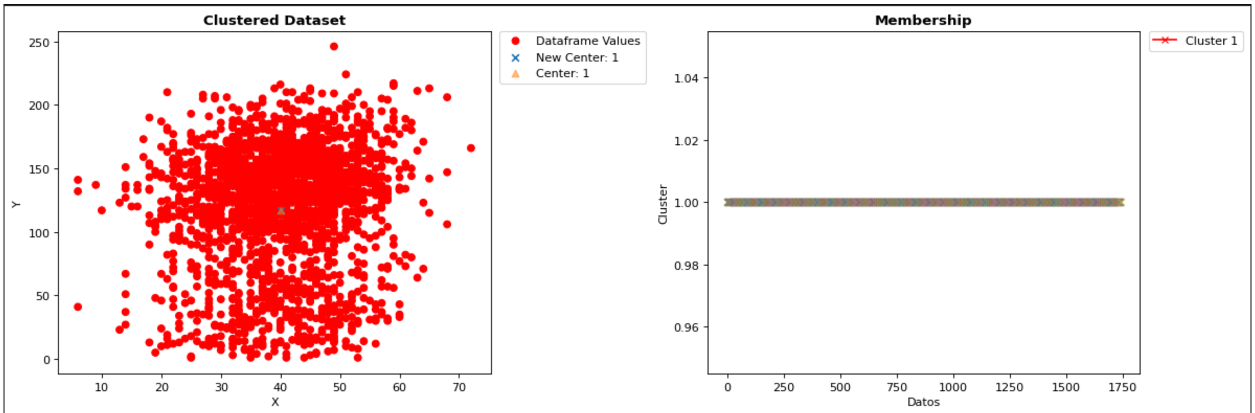


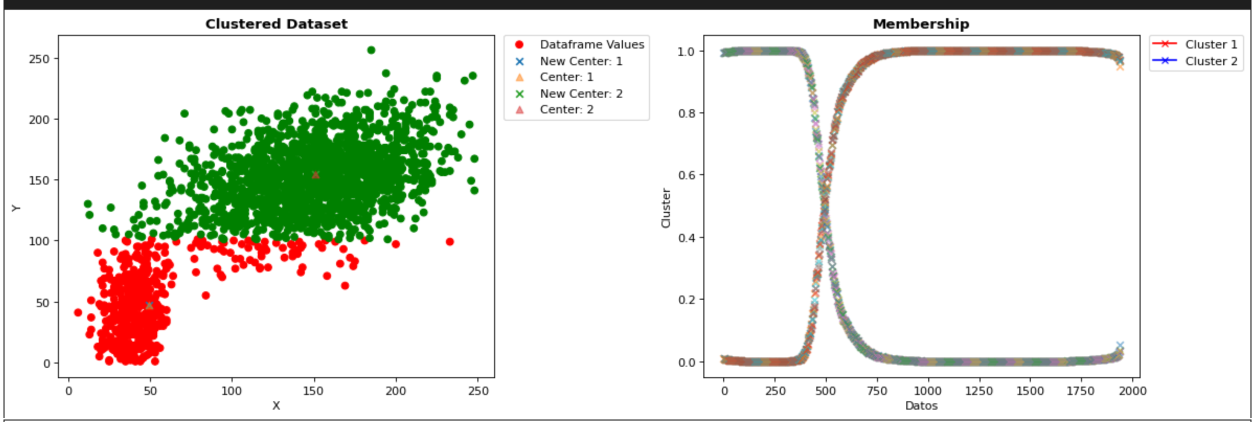

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

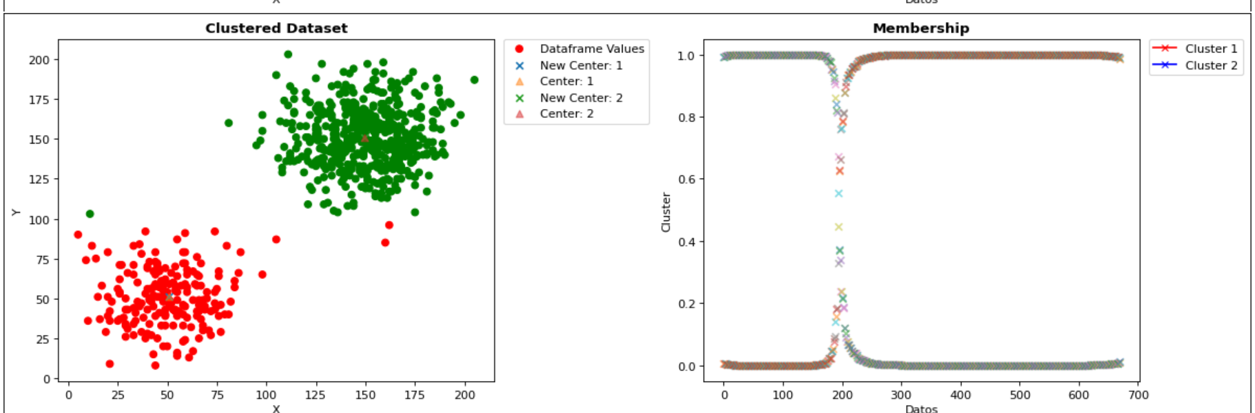

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

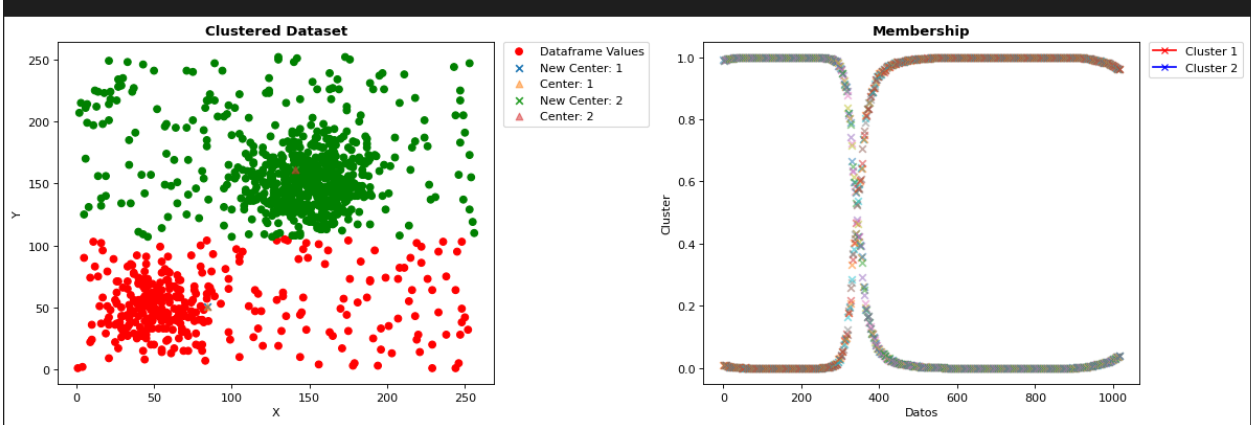

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos


Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

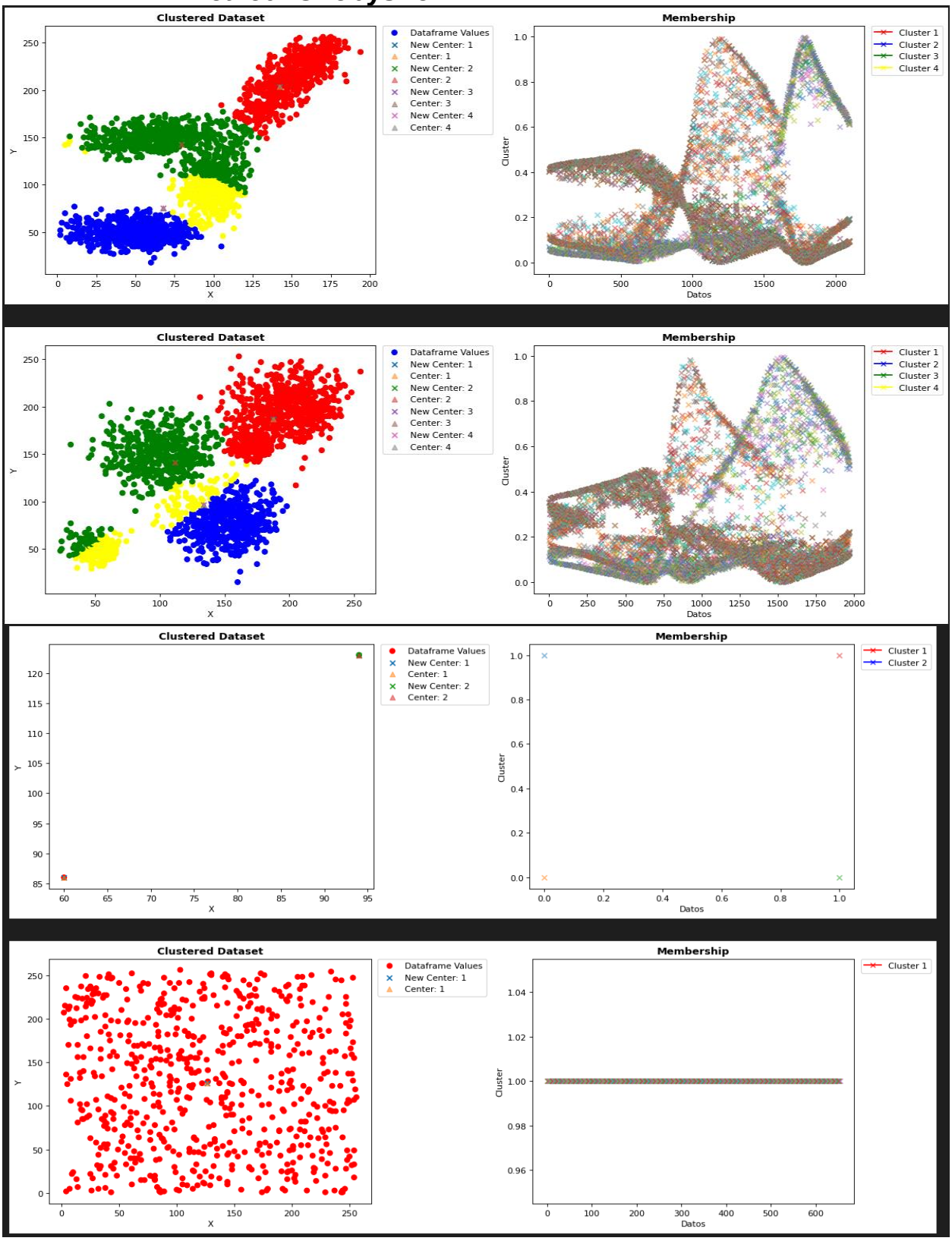

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
X
Y
Membership
Cluster 1
Datos
Cluster

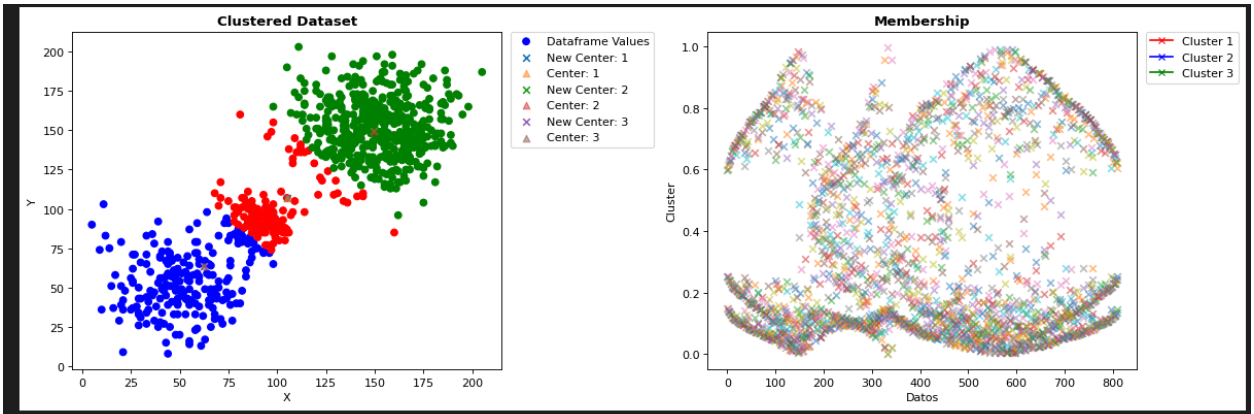

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster

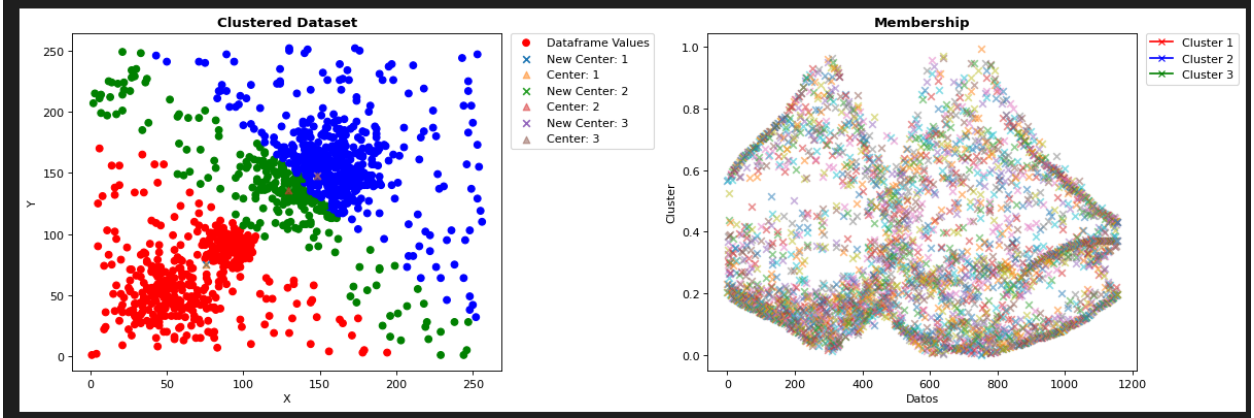

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster

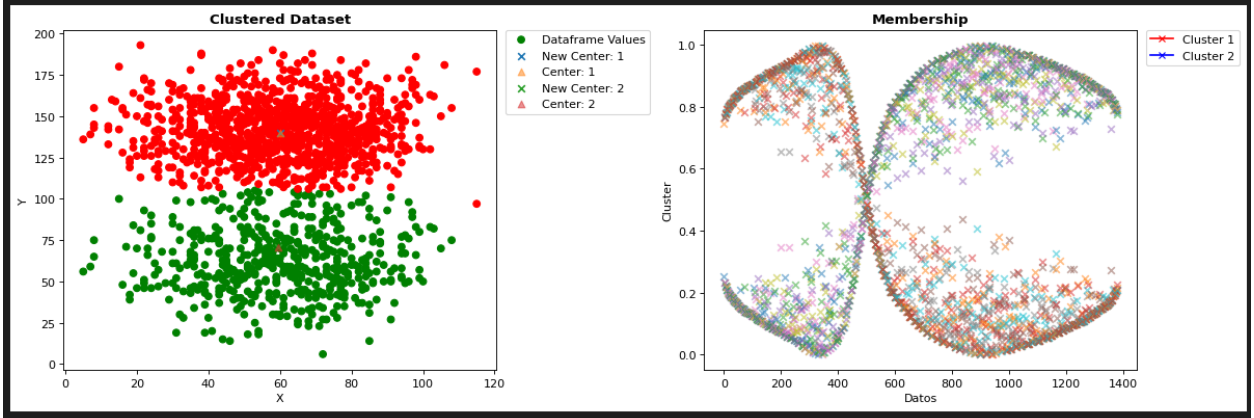

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster

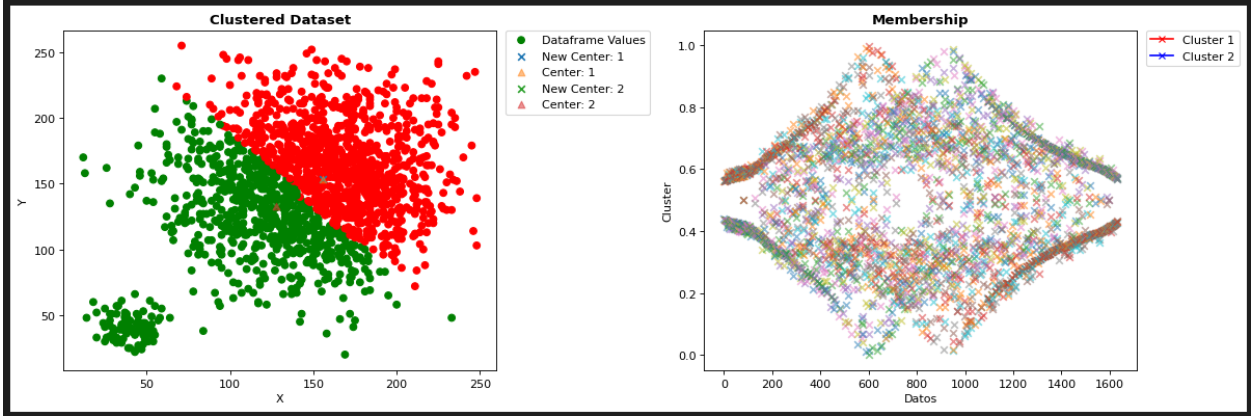
i. M = **1.5**; Initialization of Centers = **First Elements**; Distance Method: **Chebyshev**

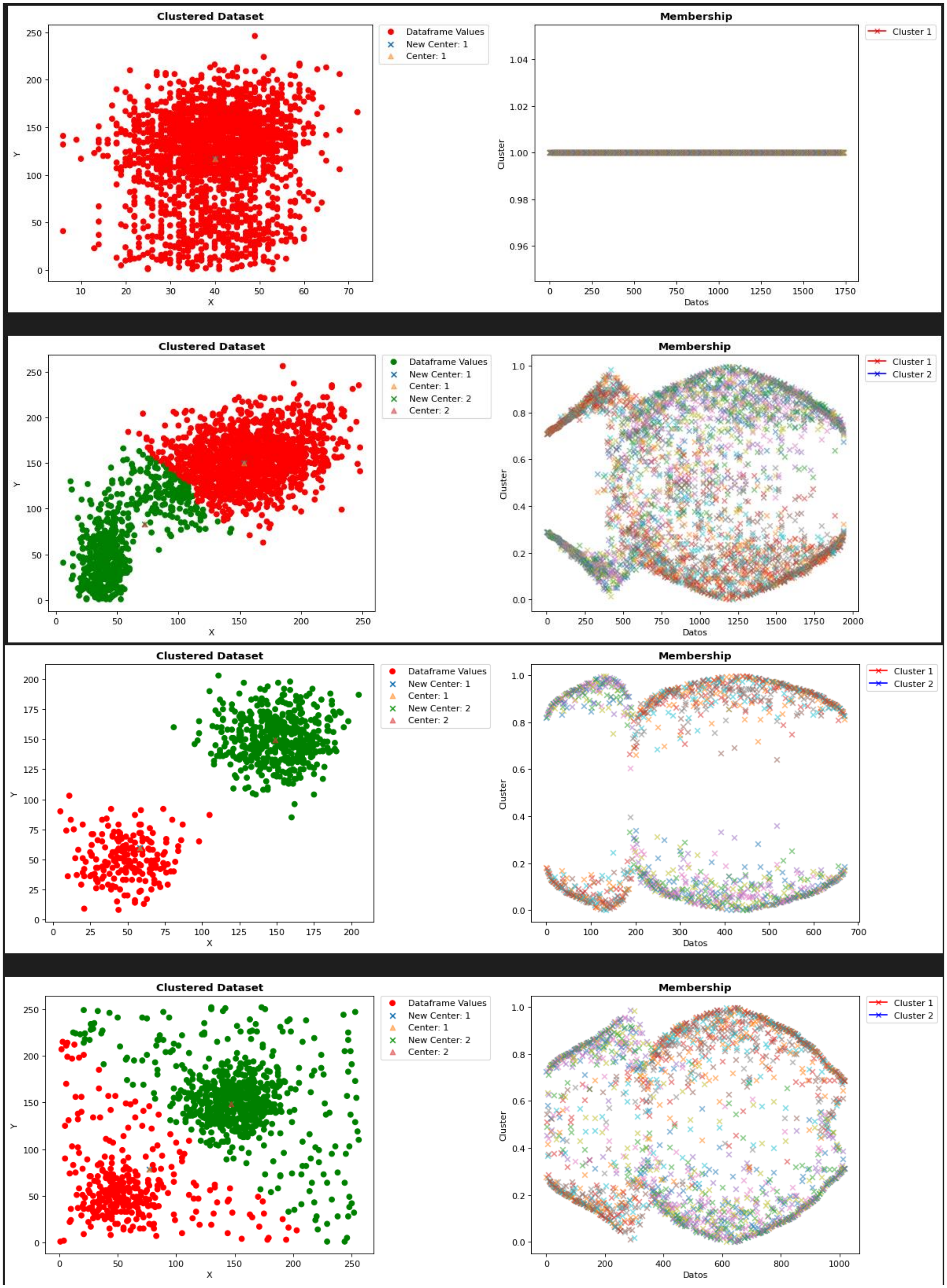



Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

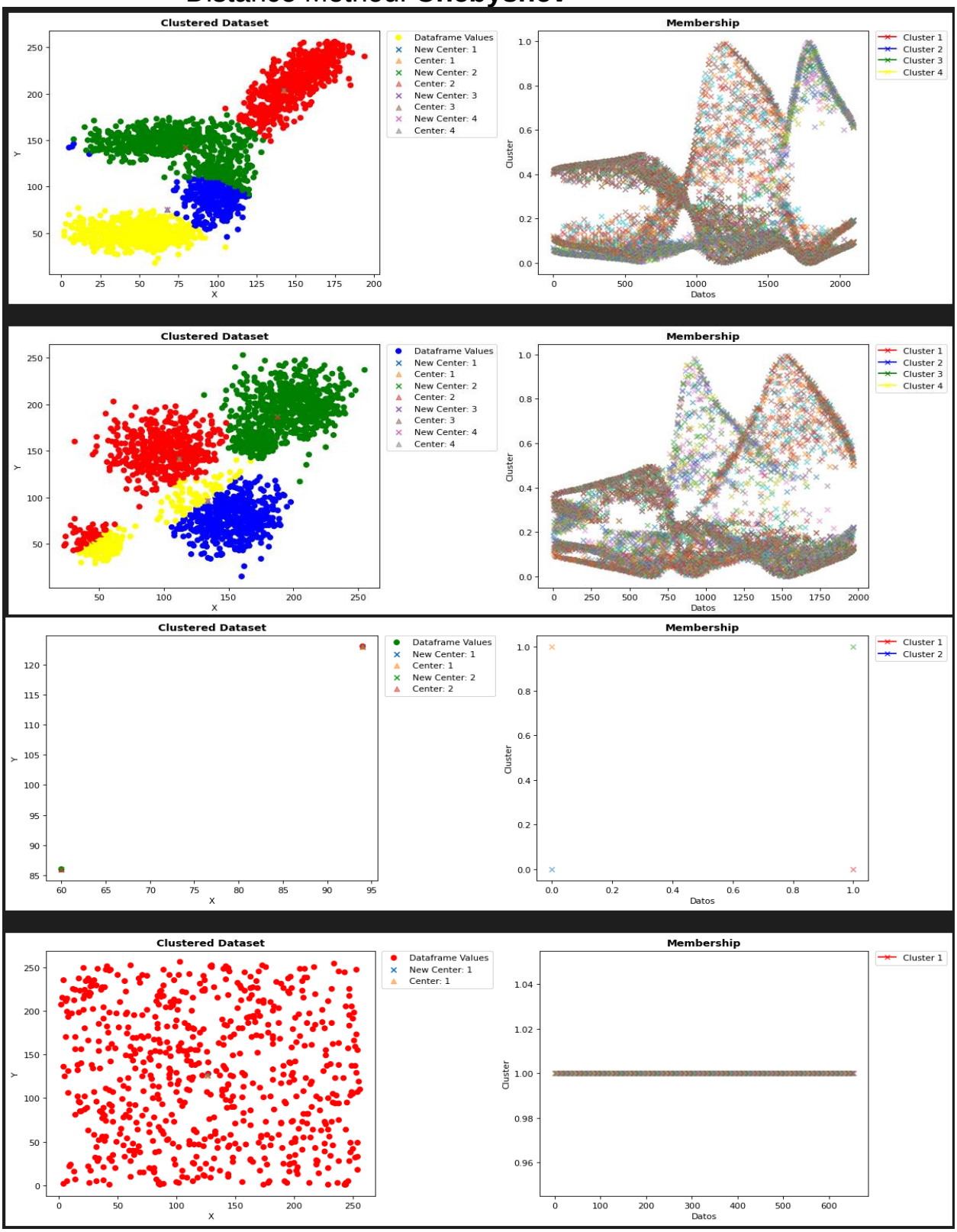

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

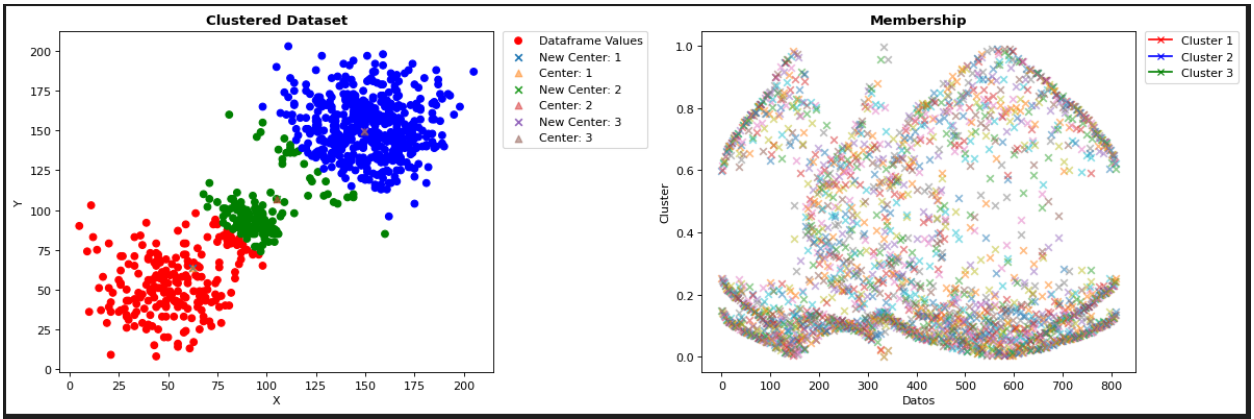

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

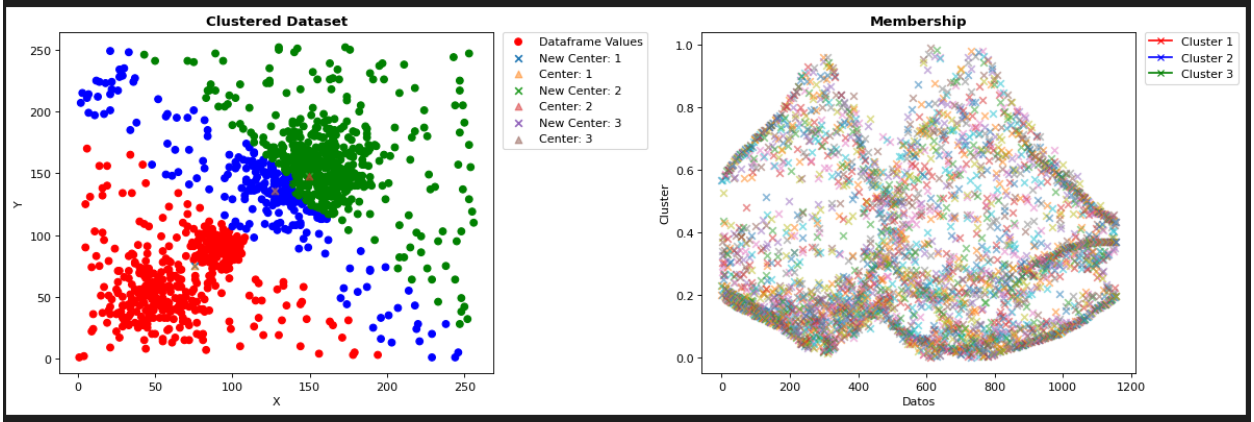

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

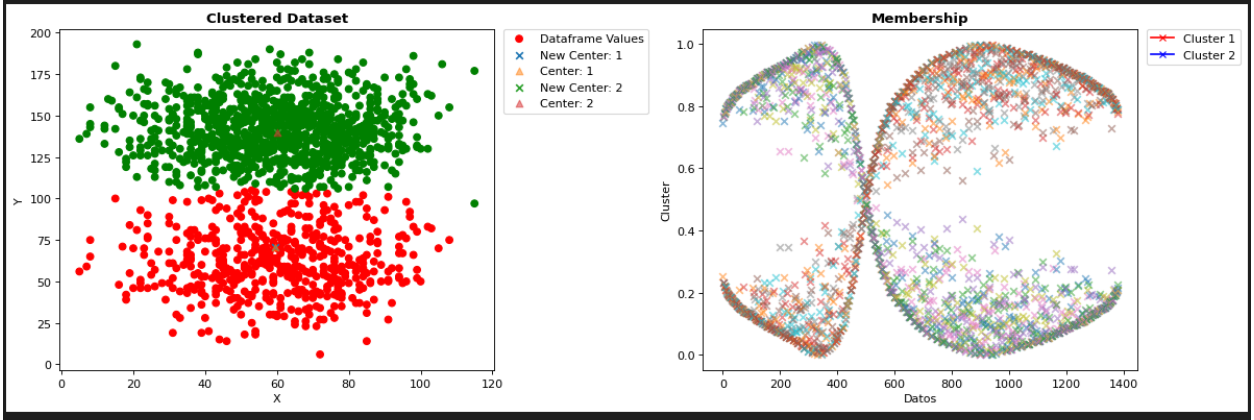

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
X
Y
Membership
Cluster 1
Cluster
Datos
Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos
Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos
Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

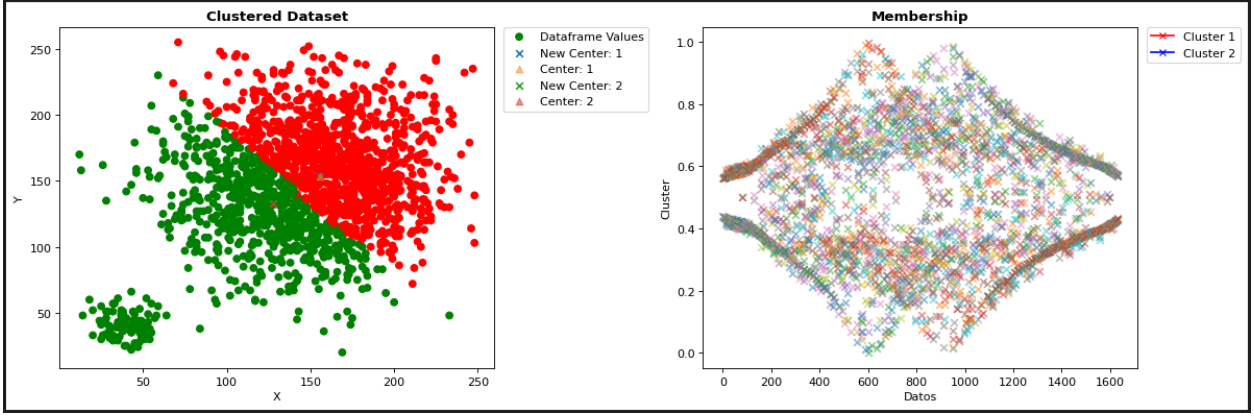
j. M = **1.5**; Initialization of Centers = **Random Elements**; Distance Method: **Chebyshev**

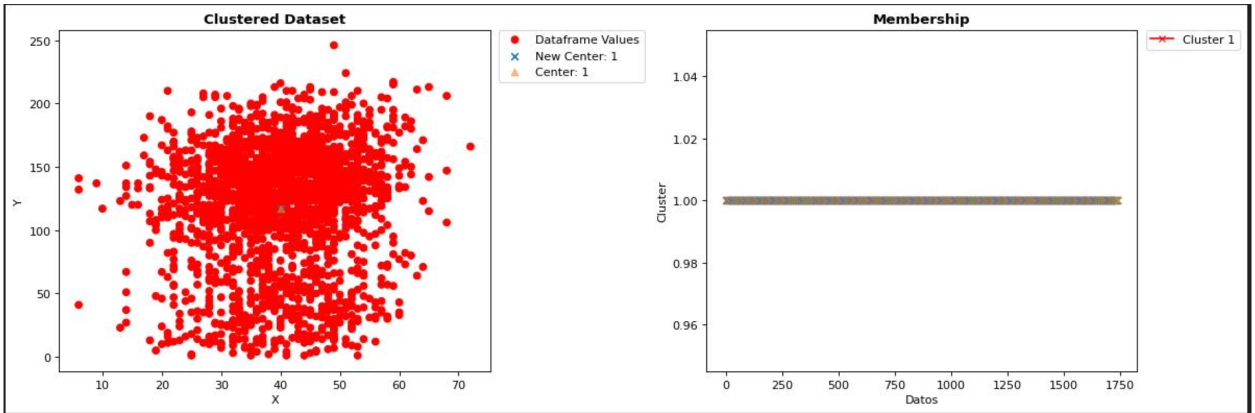


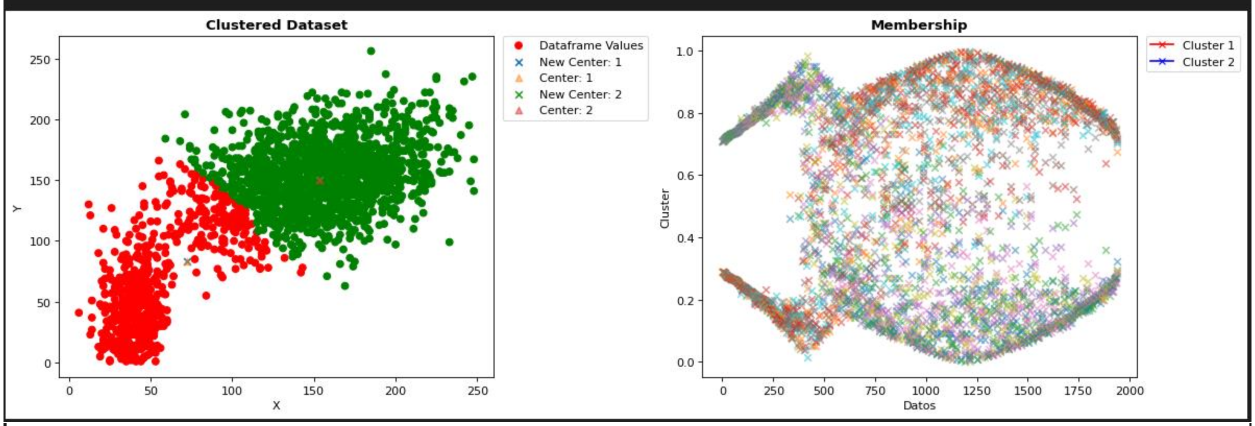

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

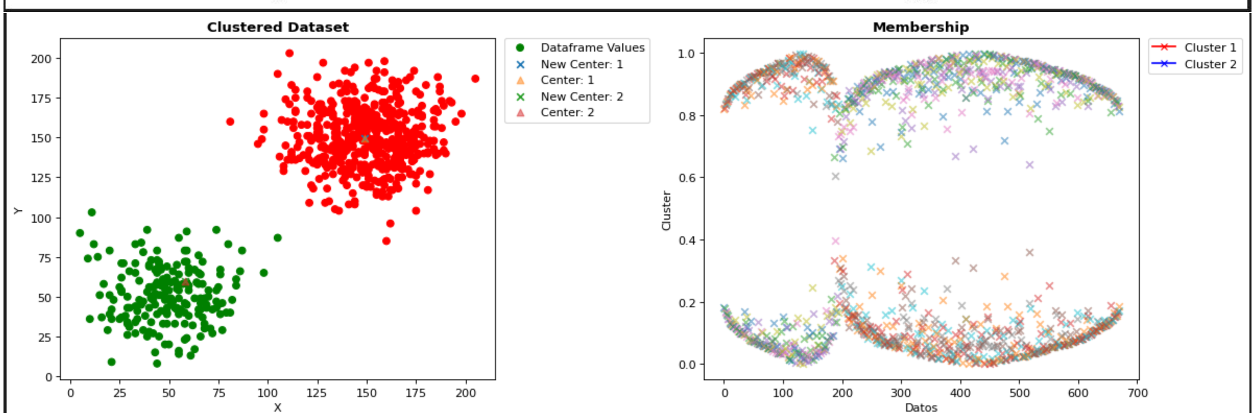

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

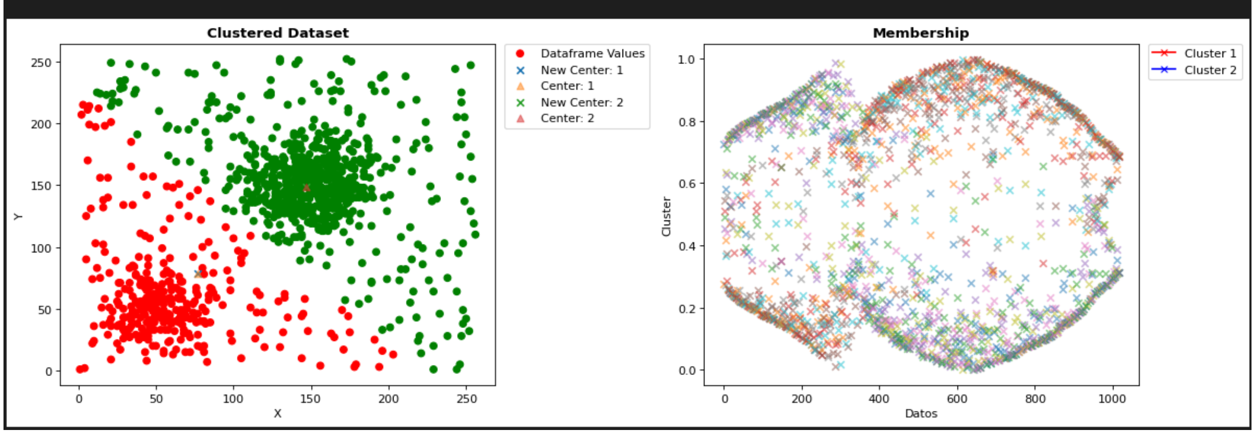

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos


Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

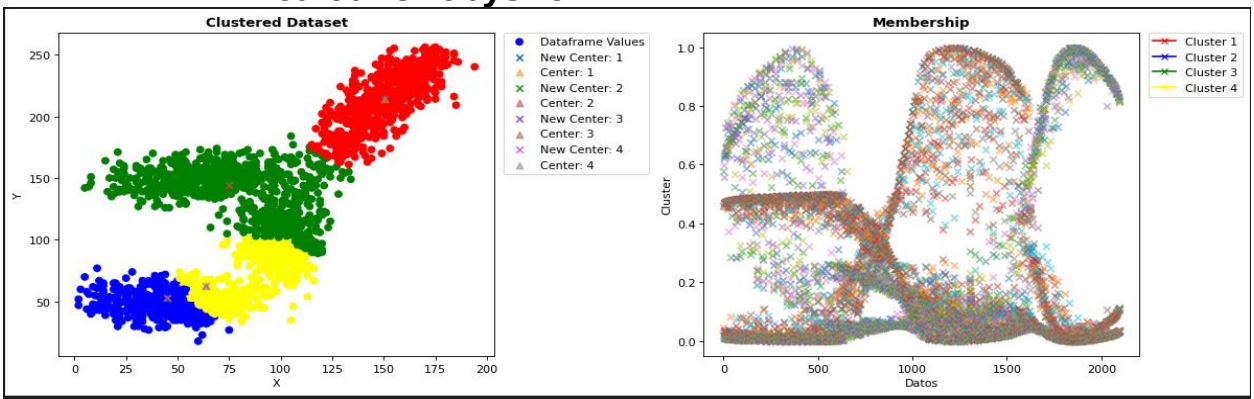

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
X
Y
Membership
Cluster 1
Cluster
Datos

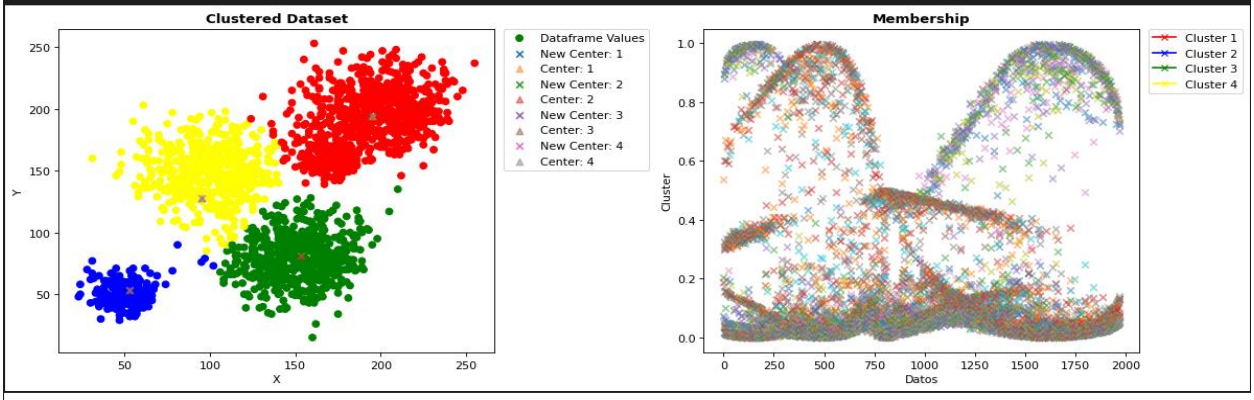

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

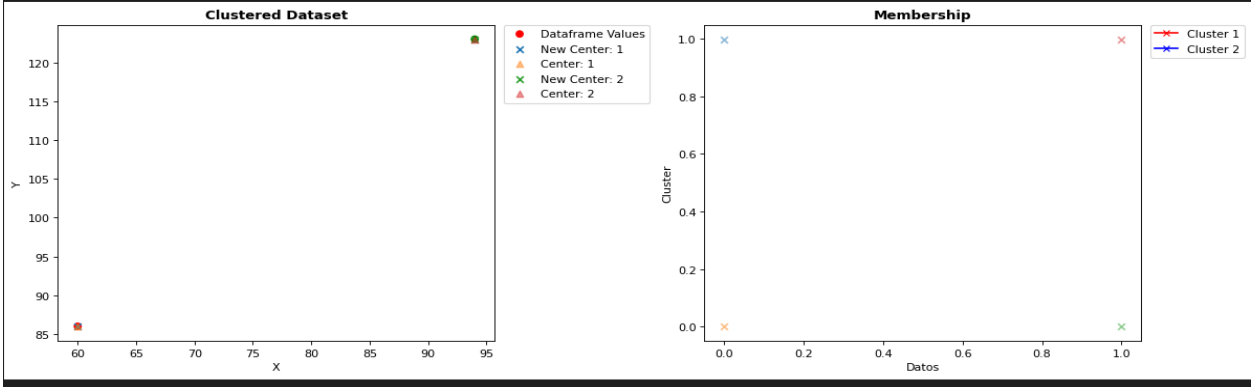

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

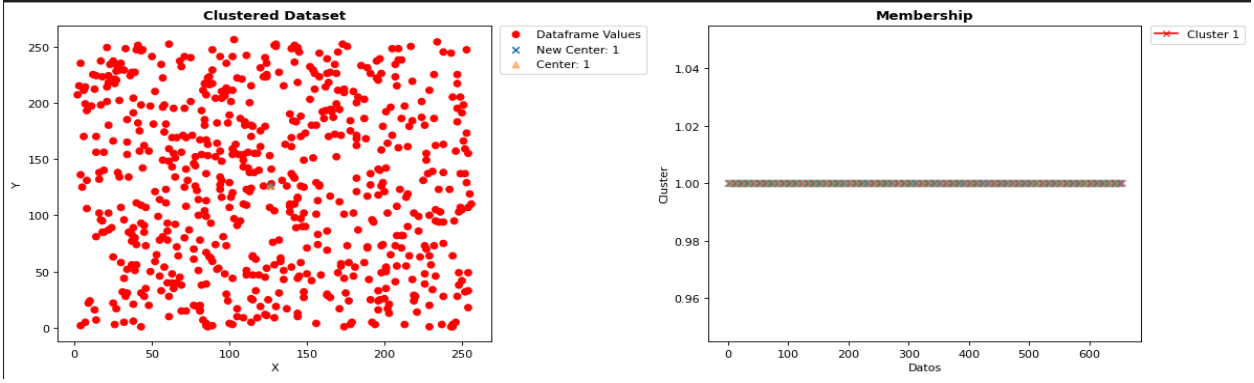

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

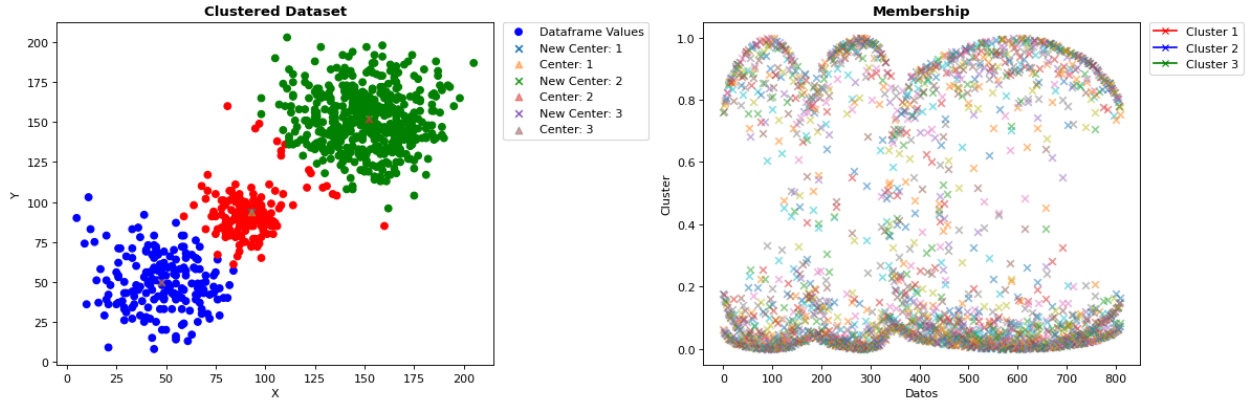
k. M = **2**; Initialization of Centers = **First Elements**; Distance Method: **Chebyshev**

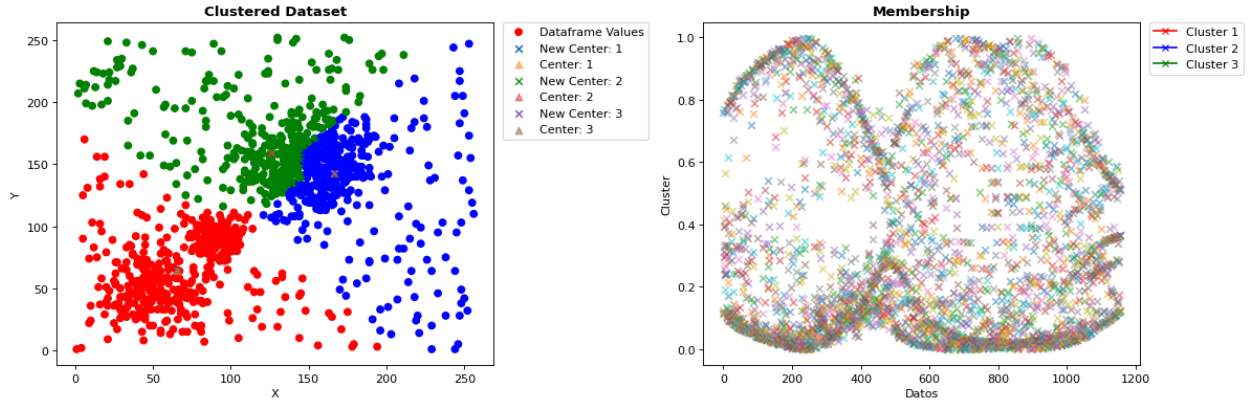


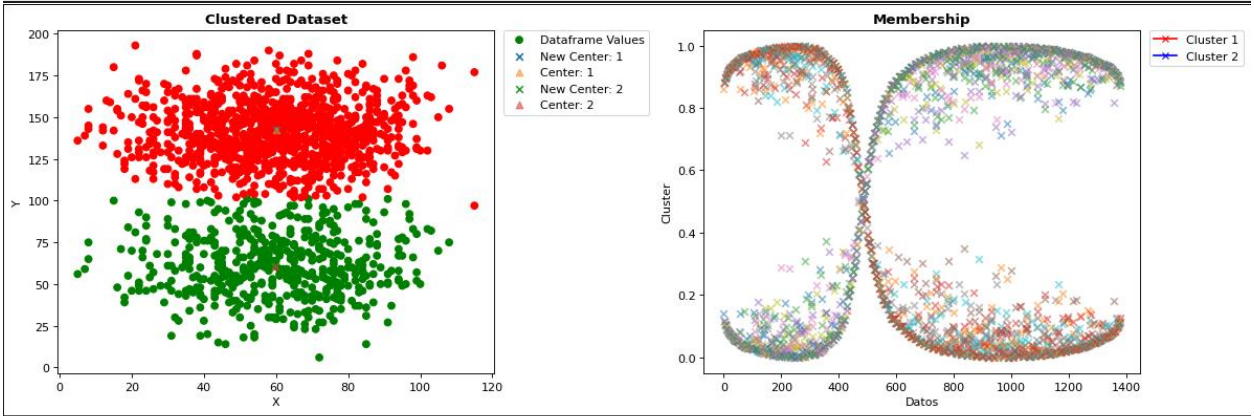


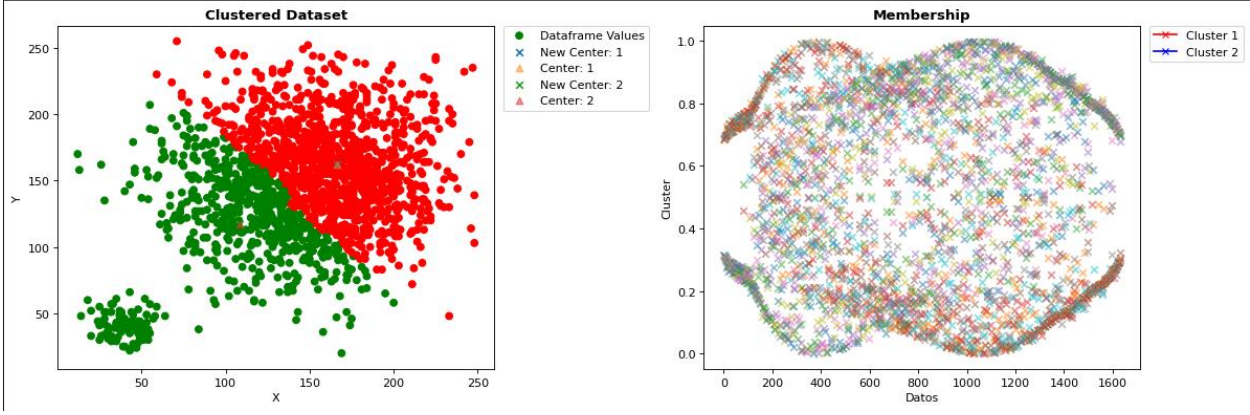


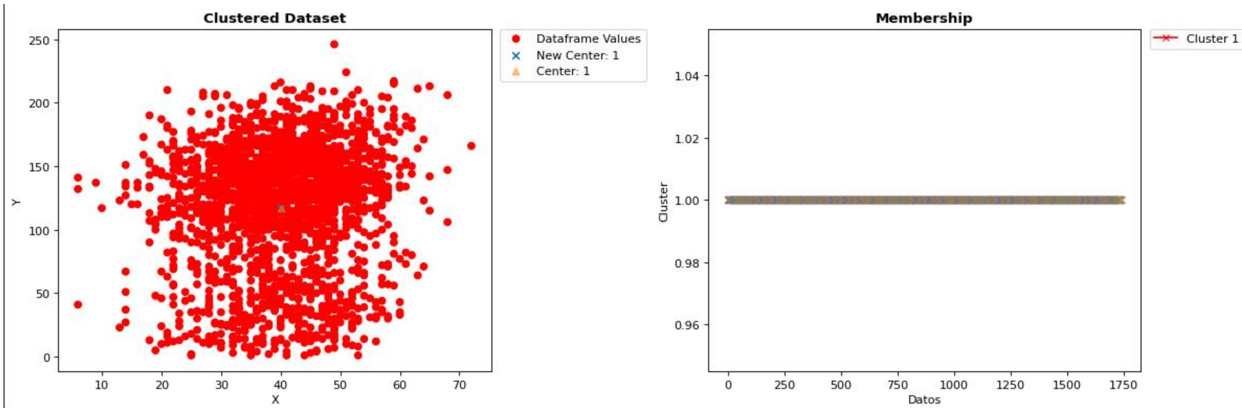


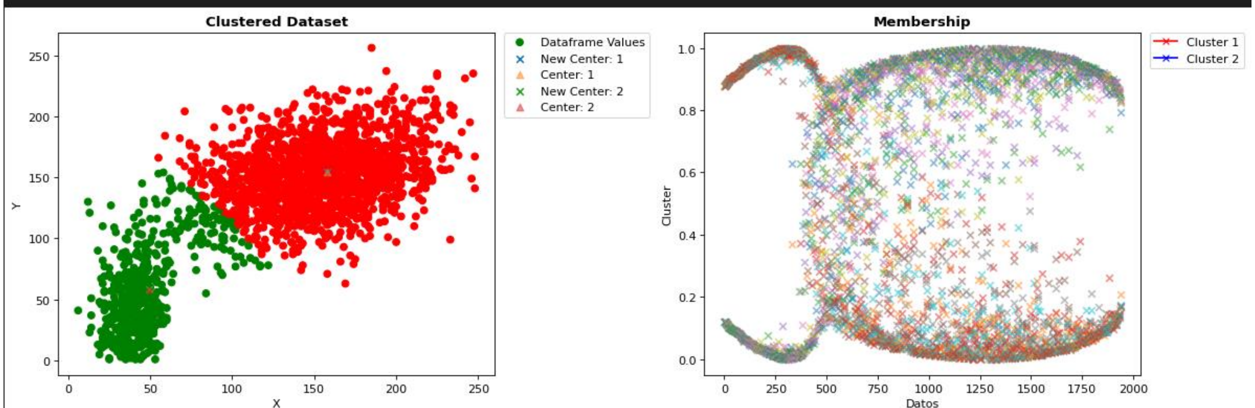

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

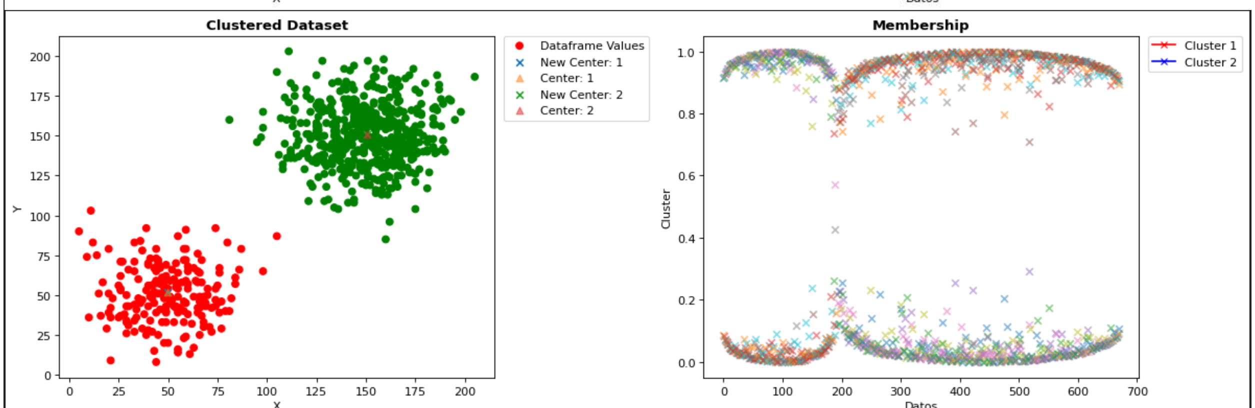

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

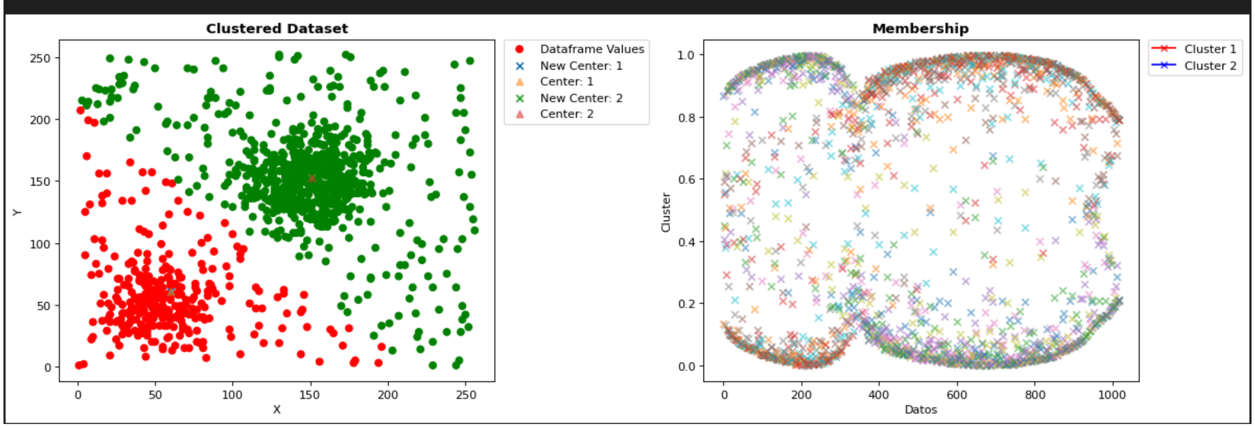

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos


Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos

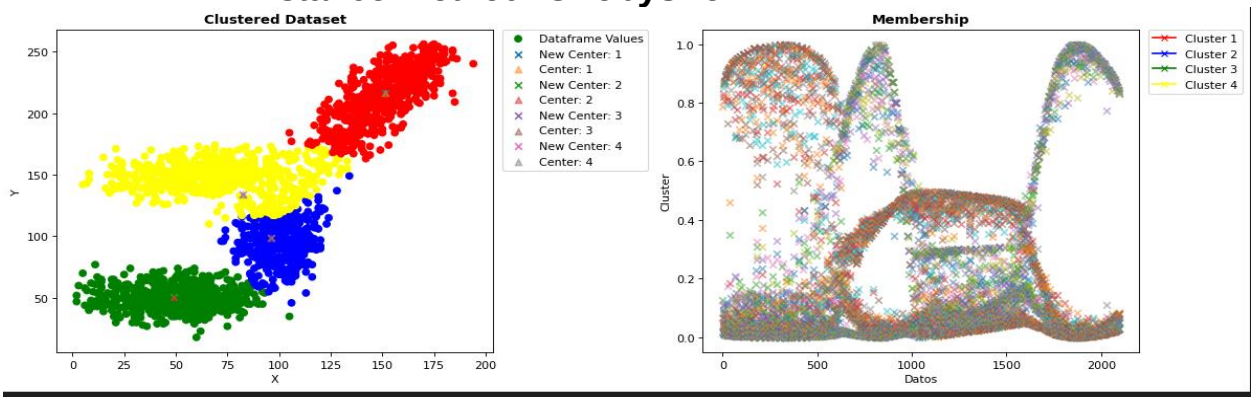

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
X
Y
Membership
Cluster 1
Datos
Cluster

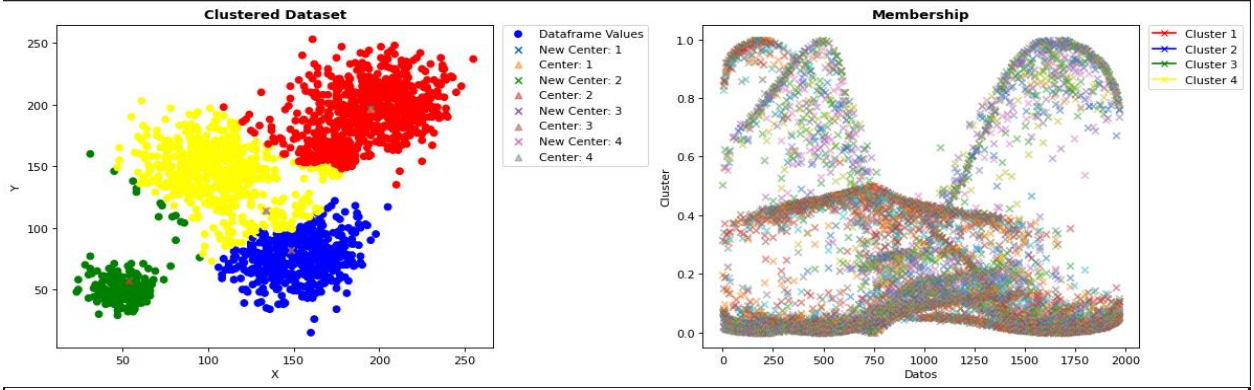

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster

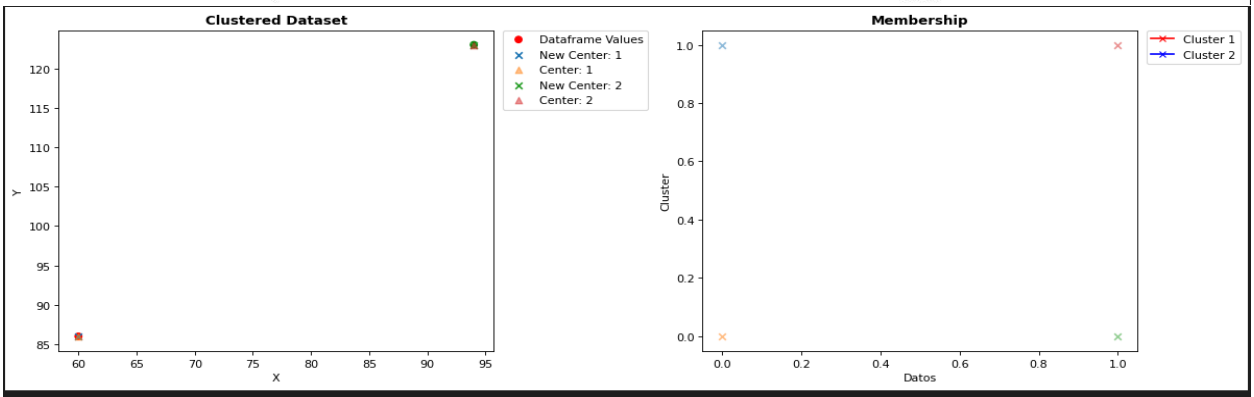

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster

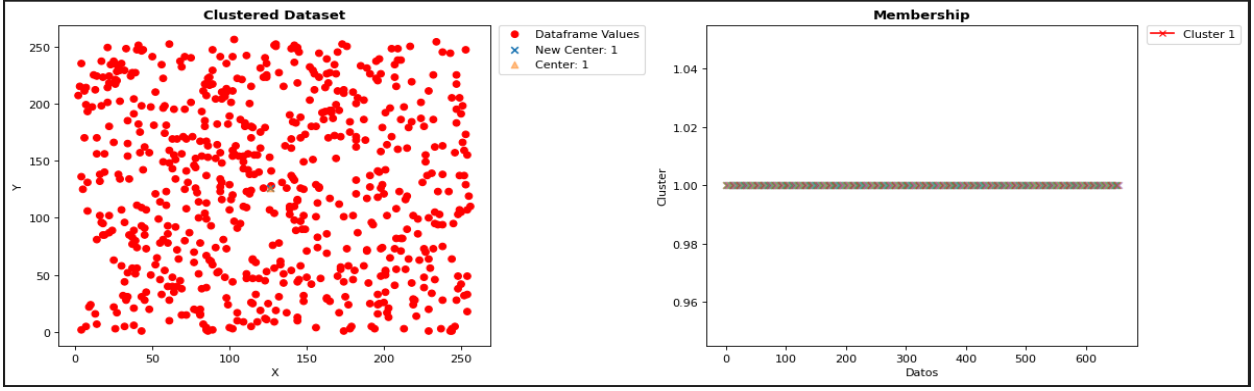

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Datos
Cluster

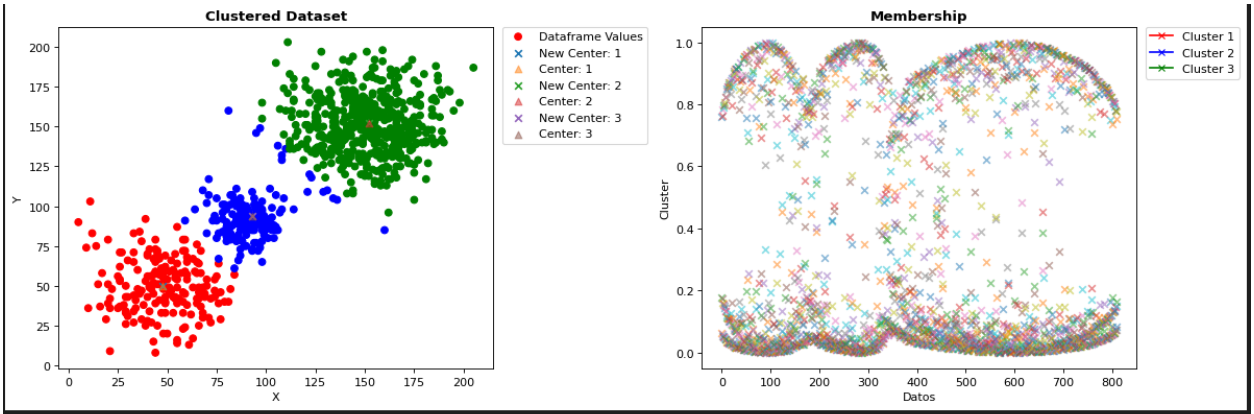
I. M = **2**; Initialization of Centers = **Random Elements**; Distance Method: **Chebyshev**

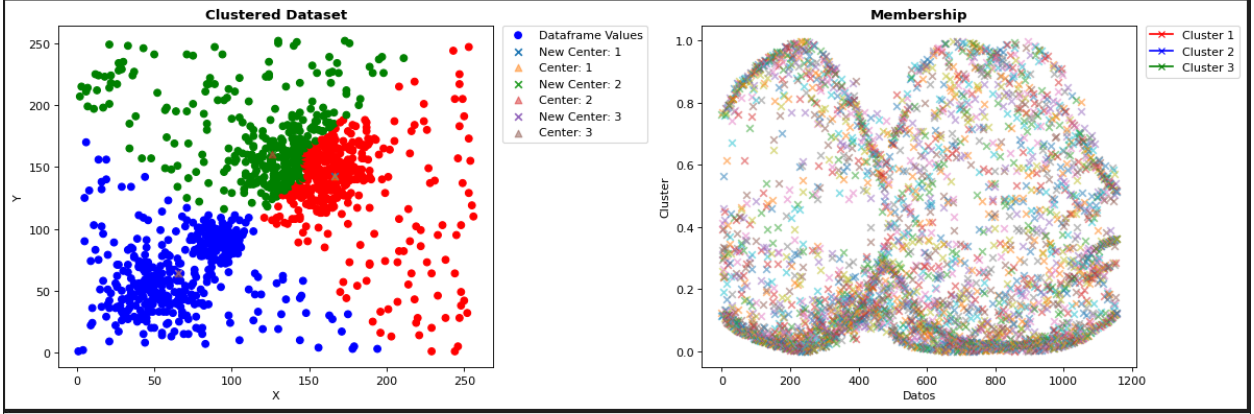


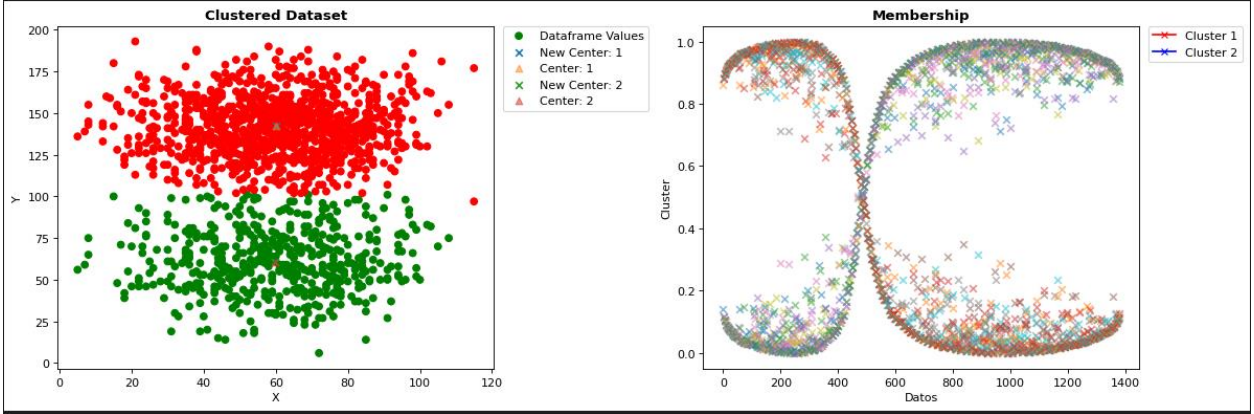


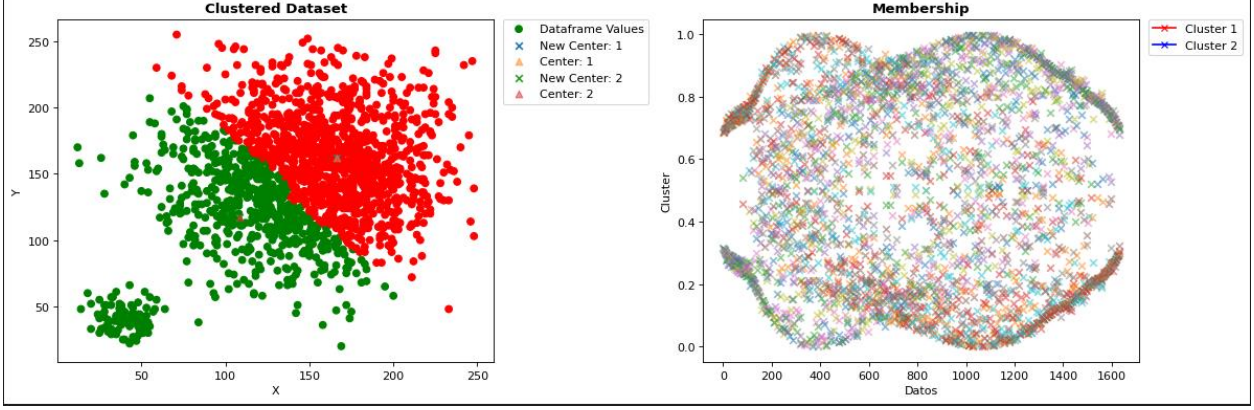


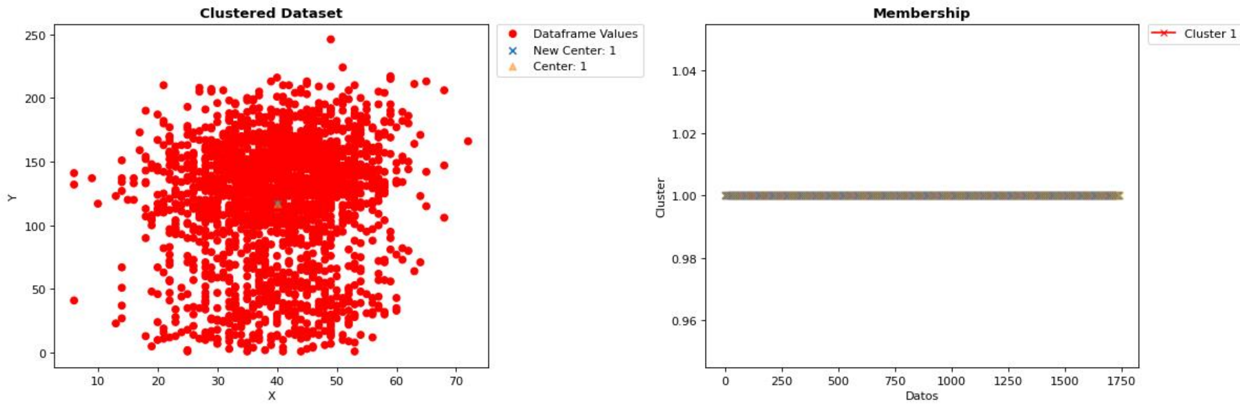


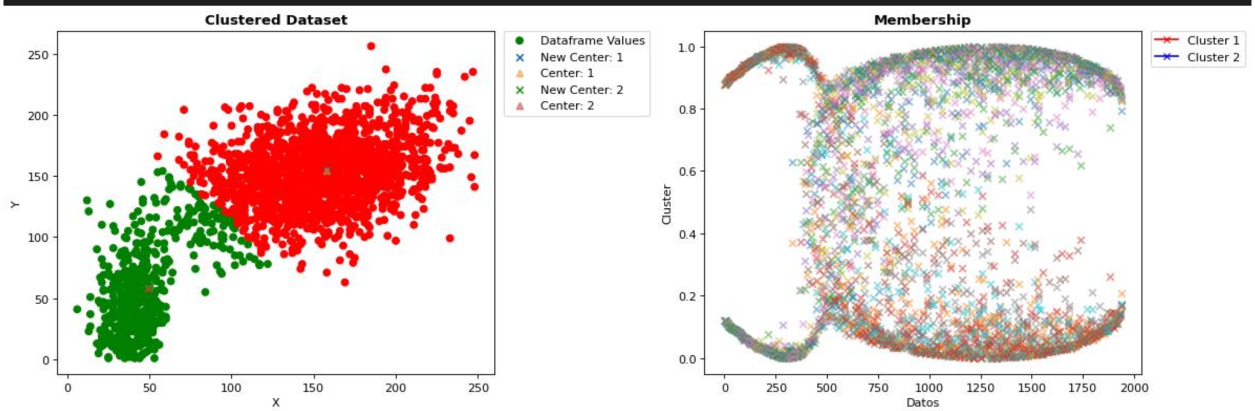

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

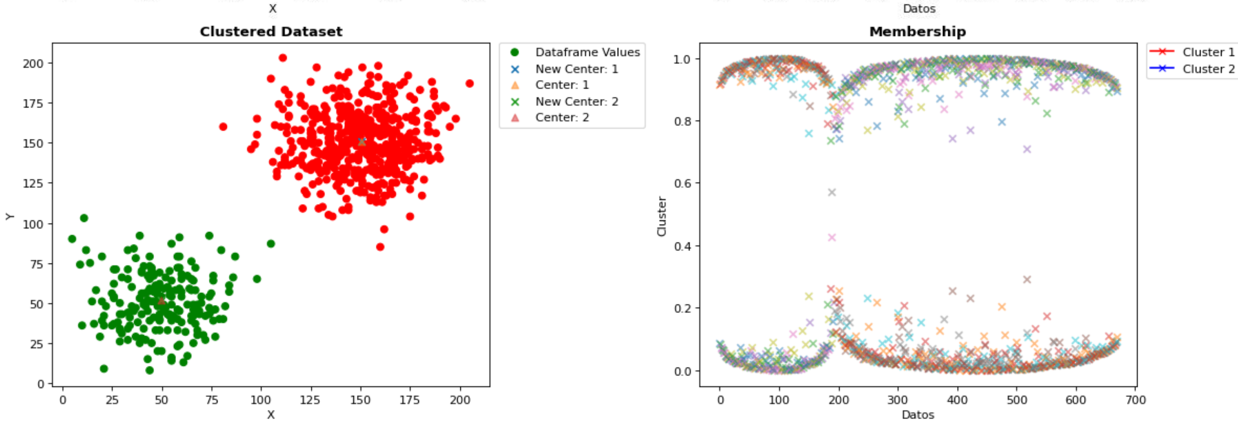

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
New Center: 3
Center: 3
X
Y
Membership
Cluster 1
Cluster 2
Cluster 3
Cluster
Datos

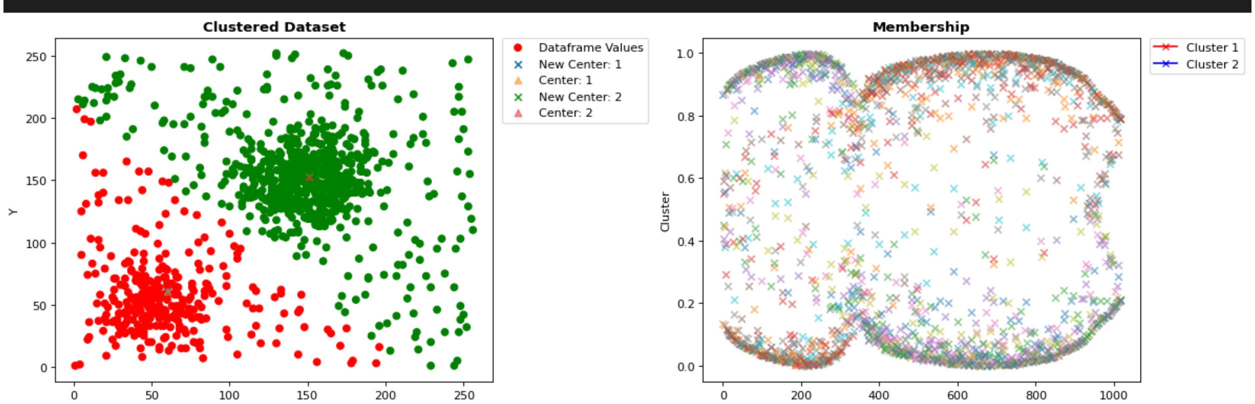

Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos


Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
X
Y
Membership
Cluster 1
Cluster 2
Cluster
Datos


Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
Y
X
Membership
Cluster 1
Cluster
Datos


Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
Y
X
Membership
Cluster 1
Cluster 2
Cluster
Datos


Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
Y
X
Membership
Cluster 1
Cluster 2
Cluster
Datos


Clustered Dataset
Dataframe Values
New Center: 1
Center: 1
New Center: 2
Center: 2
Y
X
Membership
Cluster 1
Cluster 2
Cluster
Datos

III. Table with all functions created by the program:

Function name	Description	Input	Output	Archive
Graphs	In this method I create a graph for each dataset to view its raw values.	N/A	N/A	Project_Soto.ipynb
centerSelect	This method receives the initialization that the user desires and fills a double array to the number of centers based on the proposed cluster size.	Dataframe, Cluster Size, Initialization option	Centers	
router	In this method the user sends a variable which contains the desired distance metric, and this method is in charge to select that specific metric to process and return the distance array.	Dataframe, Distance Method Option, Centers	Distance	

MembershipU	This method is responsible for determining from which group pertains each value of the dataset based on its distance and returns that calculation.	Centers, Distances, M value	Membership	
groupCenters	In this method the membership array is received and continues to calculate the new centers based on the membership provided.	Cluster Size, Membership, First Column of DF, Second Column of DF, M value	New Center, Groups	
Manual FCM	This method is the one responsible to calculate the Fuzzy C Means algorithm for each dataset.	Dataframe, Cluster Size, M value, Centers	Membership, New Centers	
Euclidean	In this method, the Euclidean formula is being performed for each value in the dataset against the selected centers.	First Column of DF, Second Column of DF, Centers	Distance	

Manhattan	In this method, the Manhattan formula is being performed for each value in the dataset against the selected centers.	First Column of DF, Second Column of DF, Centers	Distance	
Chebyshev	In this method, the Chebyshev formula is being performed for each value in the dataset against the selected centers.	First Column of DF, Second Column of DF, Centers	Distance	
Auto FCM	This method performs the FCM algorithm, except, this method repeats the process for all the datasets at once.	N/A	Memberships, New Centers	
PC	In this method the Partial Coefficient index is calculated and returned.	Membership	PC	
FS	In this method the Fukuyama-	Membership, New	FS	

	Sugeno index is calculated and returned.	Centers, Dataframe		
Ball	In this method the Ball index is calculated and returned.	Membership, New Centers, Dataframe	Ball	
Index	Its responsible to return a table of all indexes of a specific dataset that is selected by the user	Cluster Size, Membership, New Center, Dataframe, M value	Display Dataframe	
Index2	It is responsible to return a table of all indexes for every dataset	Cluster Size, Membership, New Center, Dataframe, M value	Display Dataframe	
Menu	It contains a user-friendly menu to choose a variety of options	N/A	N/A	