

Proximity Measures & Cluster Analysis

Rajesh K. Jat

2021kpad1001@iiitkota.ac.in

Statistics for Data Science, Winter (2021-22)

Dataset Used:

Shape having 8 different datasets named as S1, S2, S3, S8.

Each shape has 3 fields, the first two are data fields and the third one is the label.
(available at: <http://cs.joensuu.fi/sipu/datasets/>)

Clustering algorithms applied:

1. K-means Clustering
2. Hierarchical Clustering
3. Dbscan Clustering
4. PAM Clustering
5. Fuzzy Clustering

Cluster evaluation methods used:

1. Plots of Clusters
2. Gap statistics to know the best no of clusters in K-means
3. Silhouette Plot

Final Observations:

1. Plots of different methods on the different datasets are attached below in this document.

2. Different algorithms work differently because:
 - Kmeans works as per the selection of k value, if it is different it creates different clusters.
 - Hierarchical clustering has some initial clusters, finally they merge into higher clusters.
 - Dbscan is density based, so when the minimum points value is increased, no. of the clusters created is decreased.
 - PAM uses medoids, which are more centrally located values, so gives better results.
 - Fuzzy clustering method works on the basis of the degree of membership so many times a data point may be part of many clusters.
3. Hyperparameters like k in k means and min_pts in Dbscan are changed and plots are given.
4. Execution time of all the algorithms was calculated and then presented using a graph and it is found that on average K-means and Fuzzy clustering methods are giving minimum execution time for all the datasets.
5. Overall Dbscan gave the best performance almost in all the cases because density is a major point when shapes are drawn

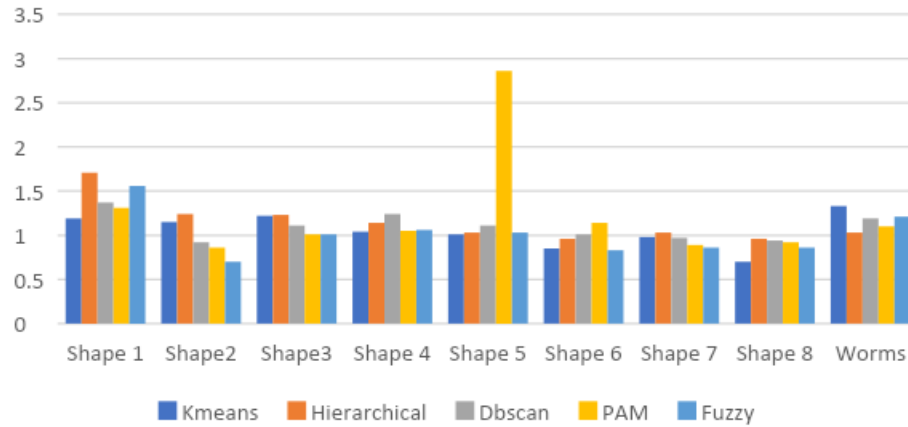
Execution time calculation was done through tic() and toc() functions. Here are the results:

Dataset	Clustering Model	Execution Time(Seconds)
Shape1	Kmeans	1.19
	Hierarchical	1.71
	Density Based	1.37
	PAM	1.31
	Fuzzy	1.56
Shape 2	Kmeans	1.15

	Hierarchical	1.24
	Density Based	0.92
	PAM	0.86
	Fuzzy	0.70
Shape 3	Kmeans	1.22
	Hierarchical	1.23
	Density Based	1.11
	PAM	1.01
	Fuzzy	1.01
Shape 4	Kmeans	1.04
	Hierarchical	1.14
	Density Based	1.24
	PAM	1.05
	Fuzzy	1.06
Shape 5	Kmeans	1.01
	Hierarchical	1.03
	Density Based	1.11
	PAM	2.86
	Fuzzy	1.03
Shape 6	Kmeans	0.85
	Hierarchical	0.96
	Density Based	1.01
	PAM	1.14
	Fuzzy	0.83
Shape 7	Kmeans	0.98
	Hierarchical	1.03
	Density Based	0.97
	PAM	0.89
	Fuzzy	0.86
Shape 8	Kmeans	0.70
	Hierarchical	0.96
	Density Based	0.94
	PAM	0.92
	Fuzzy	0.86

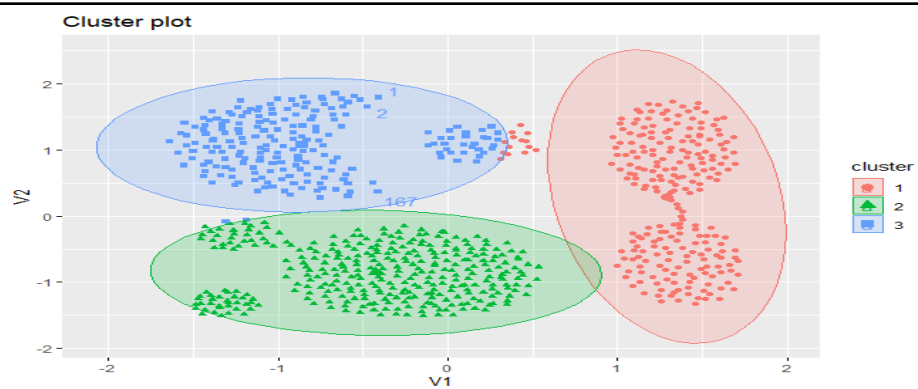
Execution Time :

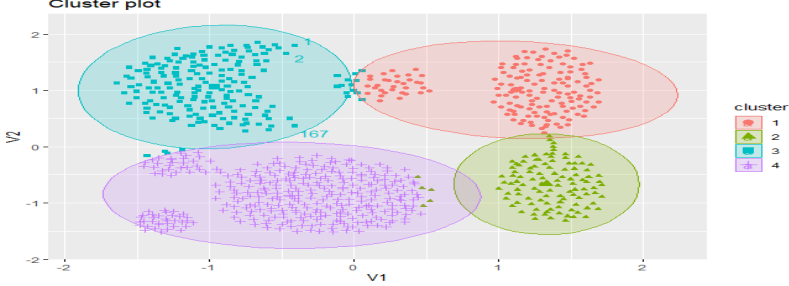
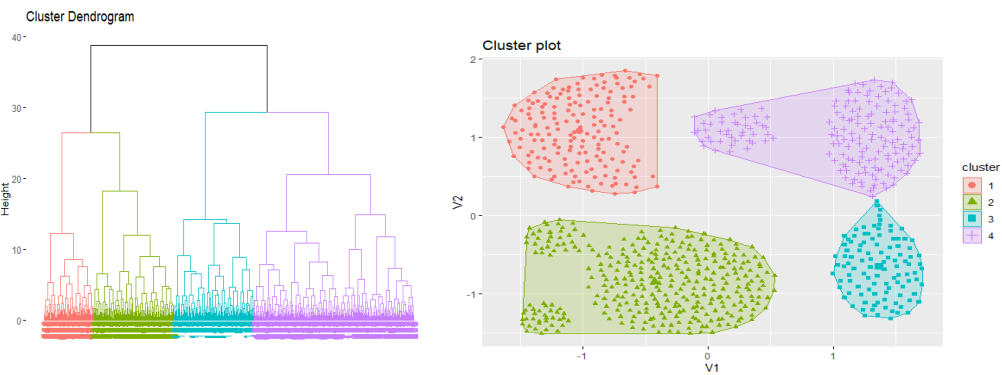
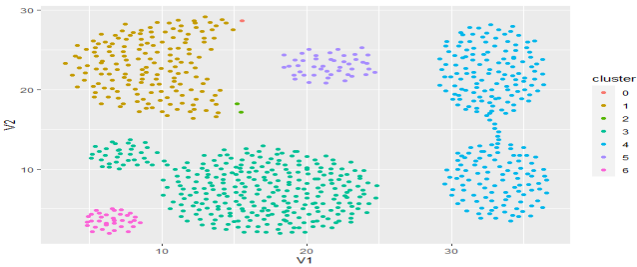
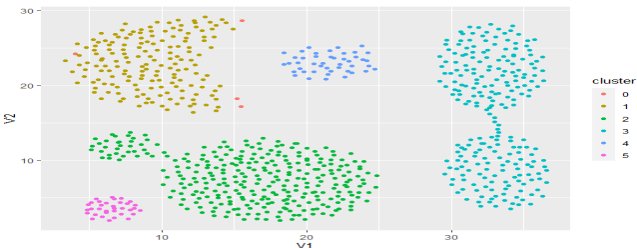
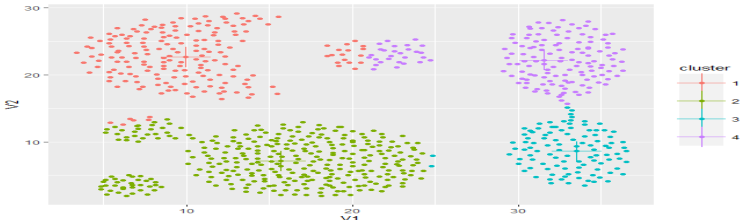
Comparison of Execution time(Seconds)
of various Clustering Algorithms

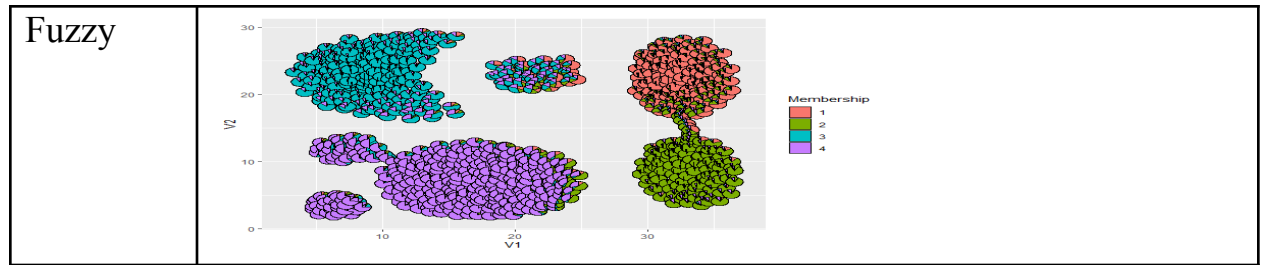


Shape 1 :

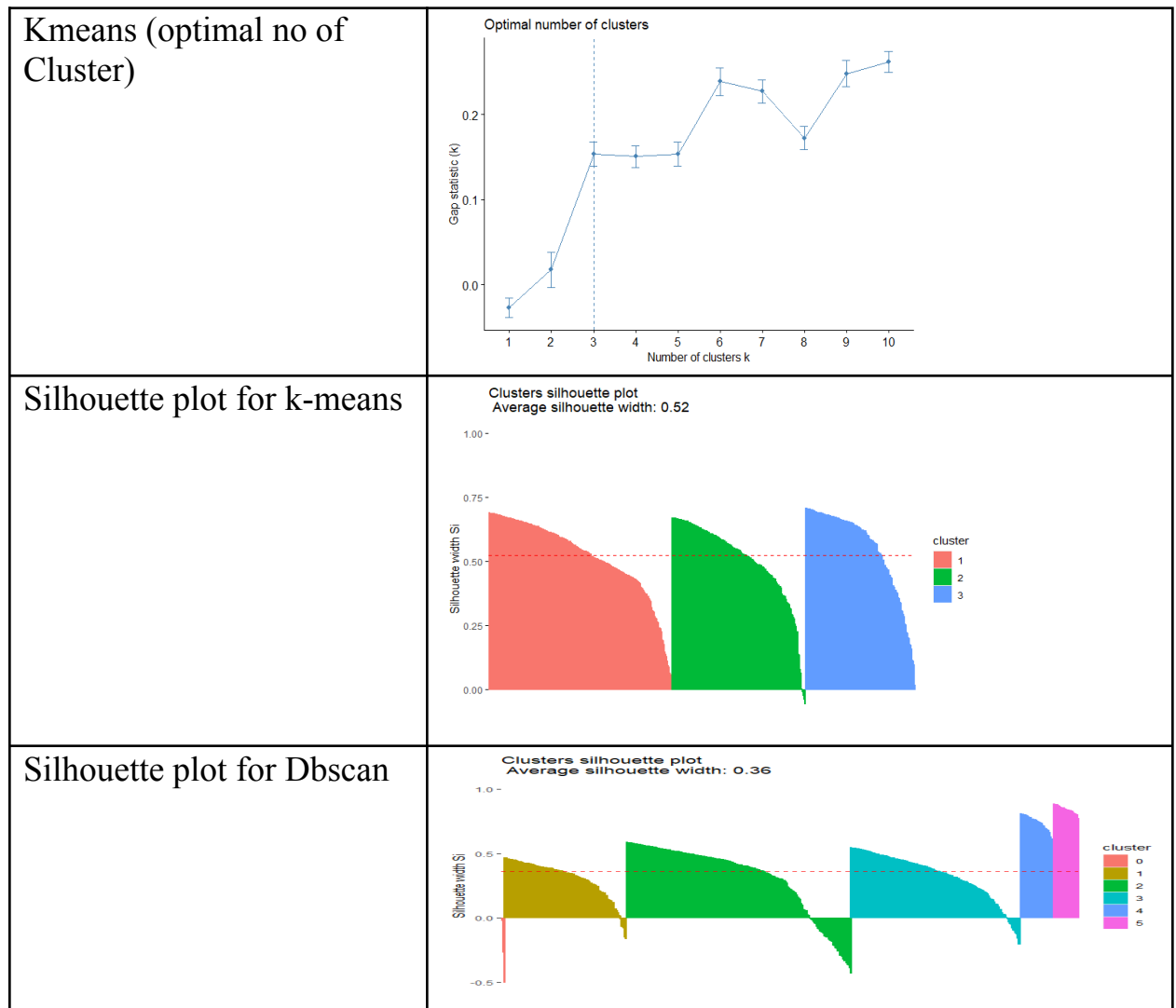
Kmeans(
K=3)



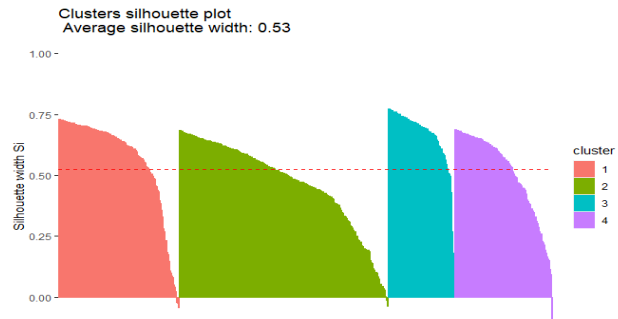
Kmeans(K=4)	 <p>Cluster plot showing 4 clusters (1, 2, 3, 4) in a 2D space (V1, V2). The clusters are represented by colored ellipses and points. The legend indicates cluster 1 (red), cluster 2 (green), cluster 3 (blue), and cluster 4 (purple).</p>
Hierarchic al	 <p>Cluster Dendrogram and Cluster plot showing hierarchical clustering. The dendrogram shows the merging of clusters, and the cluster plot shows the resulting 4 clusters (1, 2, 3, 4) in a 2D space (V1, V2). The legend indicates cluster 1 (red), cluster 2 (green), cluster 3 (blue), and cluster 4 (purple).</p>
Dbscan Minimum point=2	 <p>Cluster plot showing 6 clusters (0, 1, 2, 3, 4, 5) in a 2D space (V1, V2). The clusters are represented by colored points. The legend indicates cluster 0 (red), cluster 1 (yellow), cluster 2 (green), cluster 3 (blue), cluster 4 (cyan), and cluster 5 (magenta).</p>
Dbscan Minimum points=4	 <p>Cluster plot showing 5 clusters (0, 1, 2, 3, 4) in a 2D space (V1, V2). The clusters are represented by colored points. The legend indicates cluster 0 (red), cluster 1 (yellow), cluster 2 (green), cluster 3 (blue), and cluster 4 (cyan).</p>
PAM	 <p>Cluster plot showing 4 clusters (1, 2, 3, 4) in a 2D space (V1, V2). The clusters are represented by colored points. The legend indicates cluster 1 (red), cluster 2 (green), cluster 3 (blue), and cluster 4 (purple).</p>



Cluster Evaluation for Shape 1:

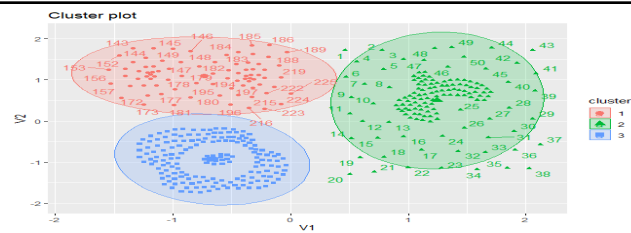


Silhouette plot for PAM

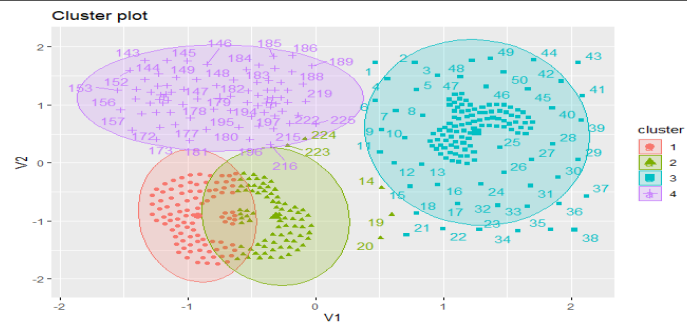


Shape2

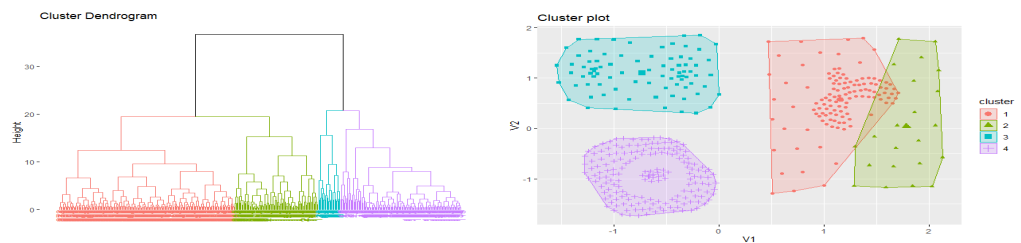
Kmeans(
K=3)

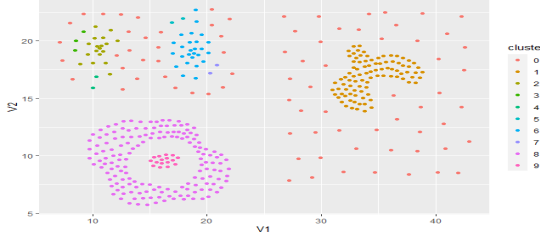
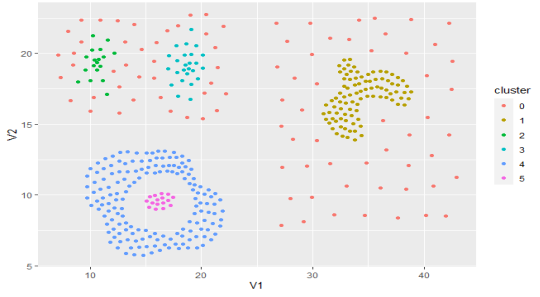
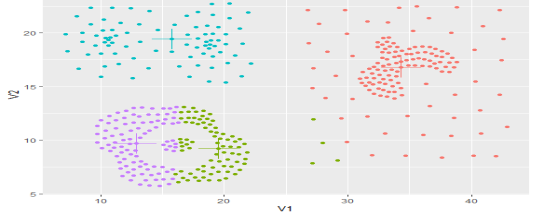
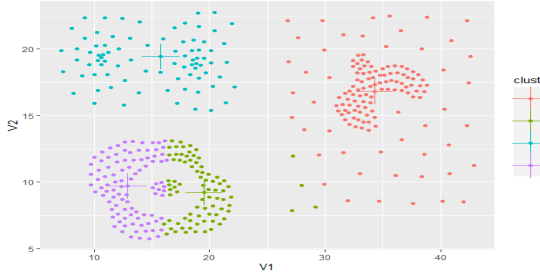


Kmeans(
K=4)

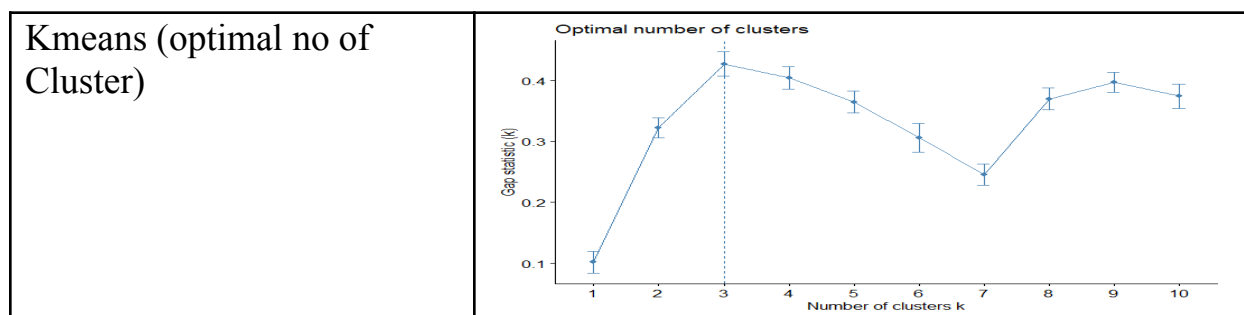


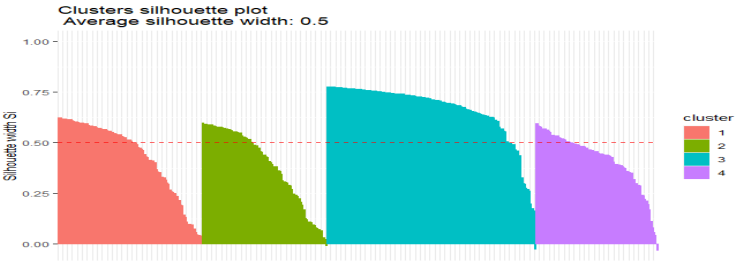
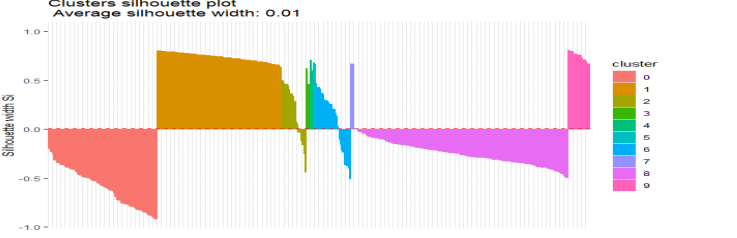
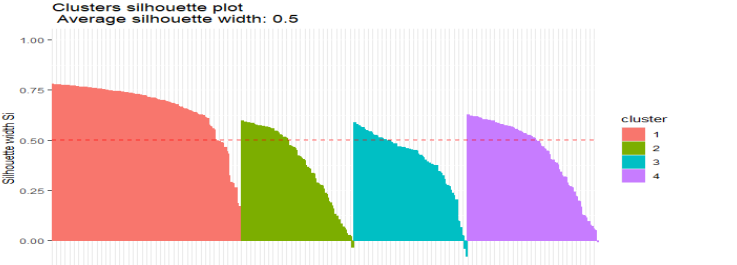
Hierarchic
al



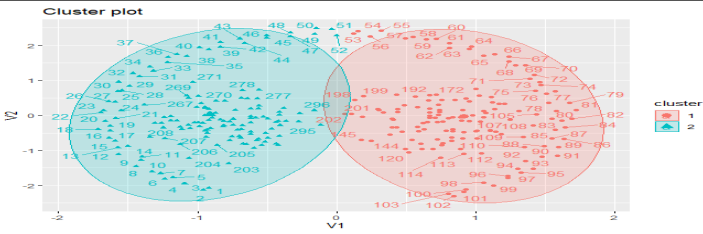
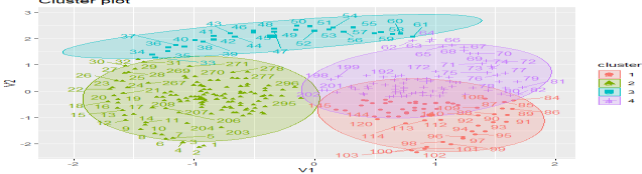
Dbscan Minimum point=2	
Dbscan Minimum points=4	
PAM	
Fuzzy	

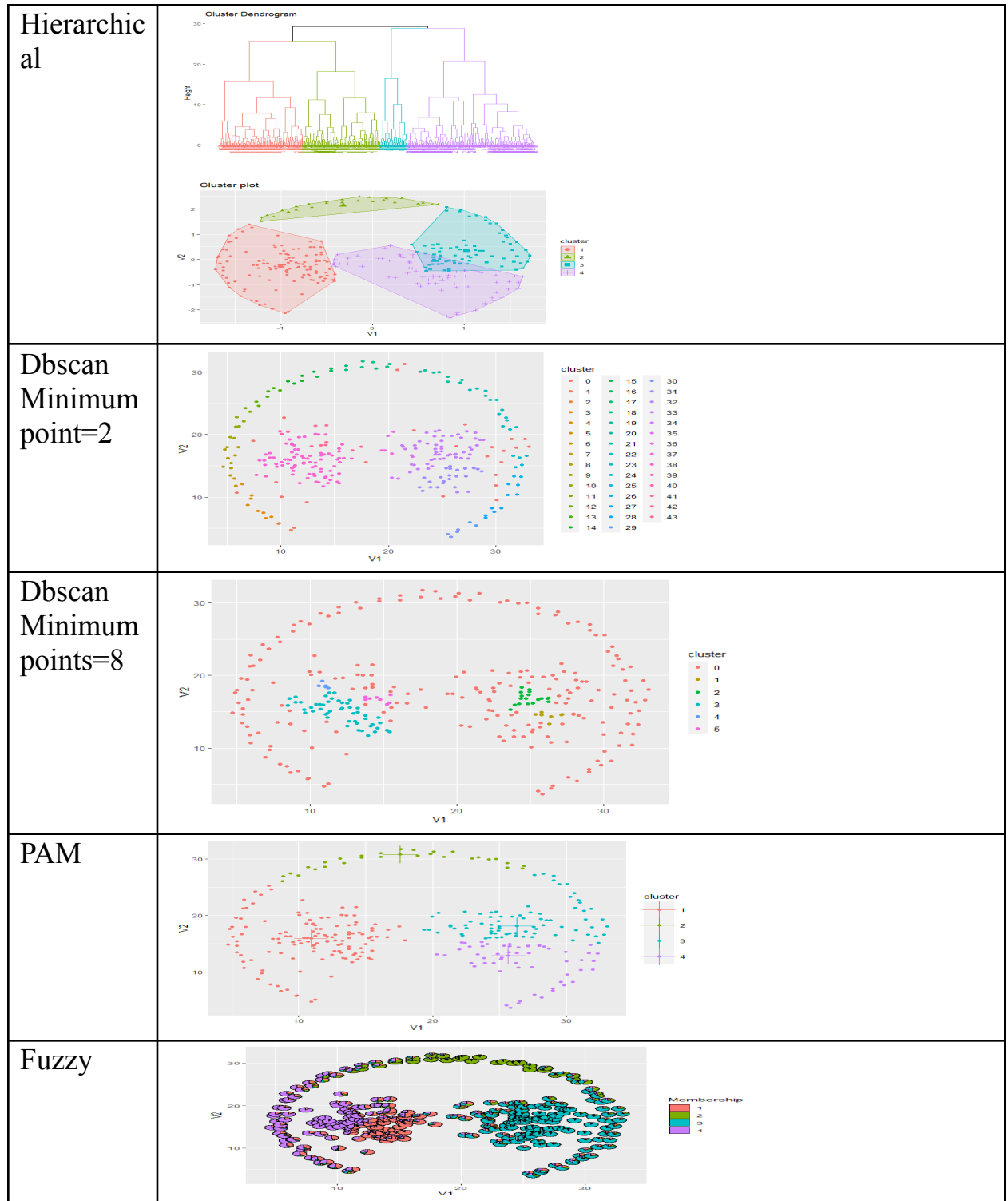
Cluster Evaluation for shape 2:



<p>Silhouette plot for k-means</p>	
<p>Silhouette plot for DbSCAN</p>	
<p>Silhouette plot for PAM</p>	

Shape 3:

<p>Kmeans(K=2)</p>	
<p>Kmeans(K=4)</p>	

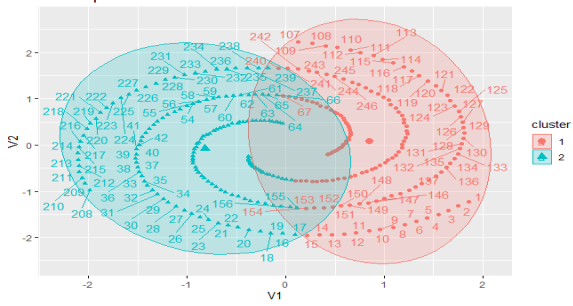
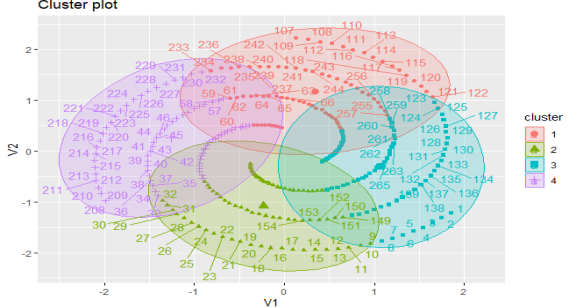
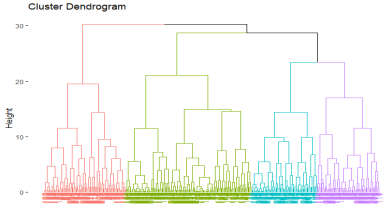
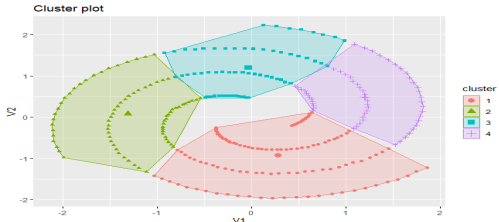
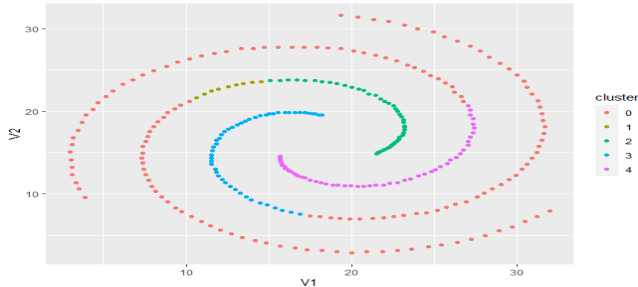
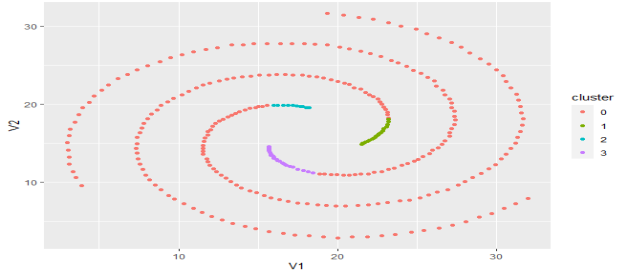


Cluster Evaluation Shape 3:

Kmeans (optimal no of Cluster)	
Silhouette plot for k-means	
Silhouette plot for Dbscan	
Silhouette plot for PAM	

Shape 4 :

Kmeans(K=2)	
-----------------	--

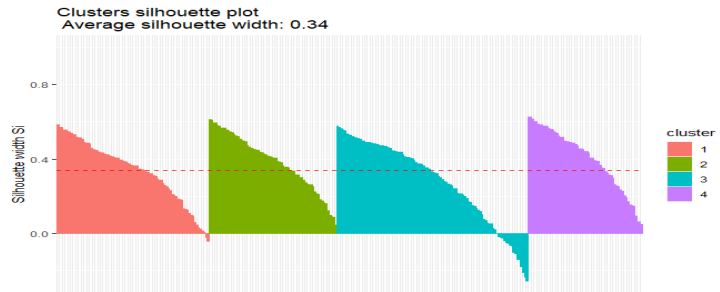
	<p>Cluster plot</p>  <p>A scatter plot showing two clusters of data points in a 2D space (V1 vs V2). The points are numbered. Cluster 1 (red) is on the right, and Cluster 2 (blue) is on the left. The axes range from -2 to 2.</p>
Kmeans(K=4)	<p>Cluster plot</p>  <p>A scatter plot showing four clusters of data points in a 2D space (V1 vs V2). The points are numbered. The clusters are colored red, blue, green, and purple. The axes range from -2 to 2.</p>
Hierarchical	<p>Cluster Dendrogram</p>  <p>A dendrogram showing the hierarchical clustering of data points. The y-axis is labeled 'Height' and ranges from 0 to 30. The x-axis represents the data points, which are grouped into four main clusters colored red, blue, green, and purple.</p> <p>Cluster plot</p>  <p>A scatter plot showing four clusters of data points in a 2D space (V1 vs V2). The points are numbered. The clusters are colored red, blue, green, and purple. The axes range from -2 to 2.</p>
Dbscan Minimum point=4	 <p>A scatter plot showing five clusters of data points in a 2D space (V1 vs V2). The points are numbered. The clusters are colored red, blue, green, purple, and yellow. The axes range from 10 to 30.</p>
Dbscan Minimum points=8	 <p>A scatter plot showing four clusters of data points in a 2D space (V1 vs V2). The points are numbered. The clusters are colored red, blue, green, and purple. The axes range from 10 to 30.</p>

PAM	
Fuzzy	

Cluster Evaluation for Shape 4.

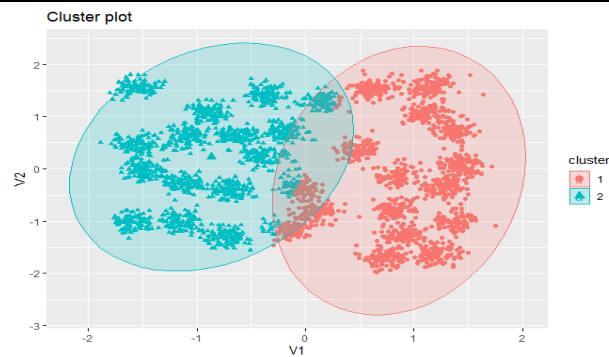
Kmeans (optimal no of Cluster)	
Silhouette plot for k-means	
Silhouette plot for Dbscan	

Silhouette plot for PAM

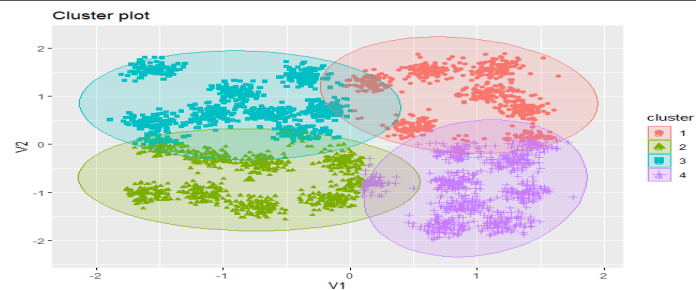


Shape 5 :

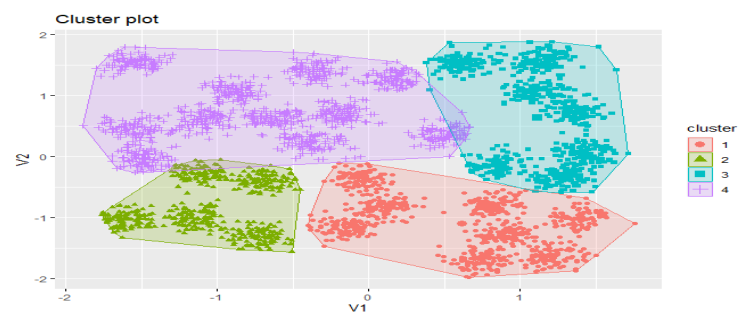
Kmeans(
K=2)



Kmeans(
K=4)

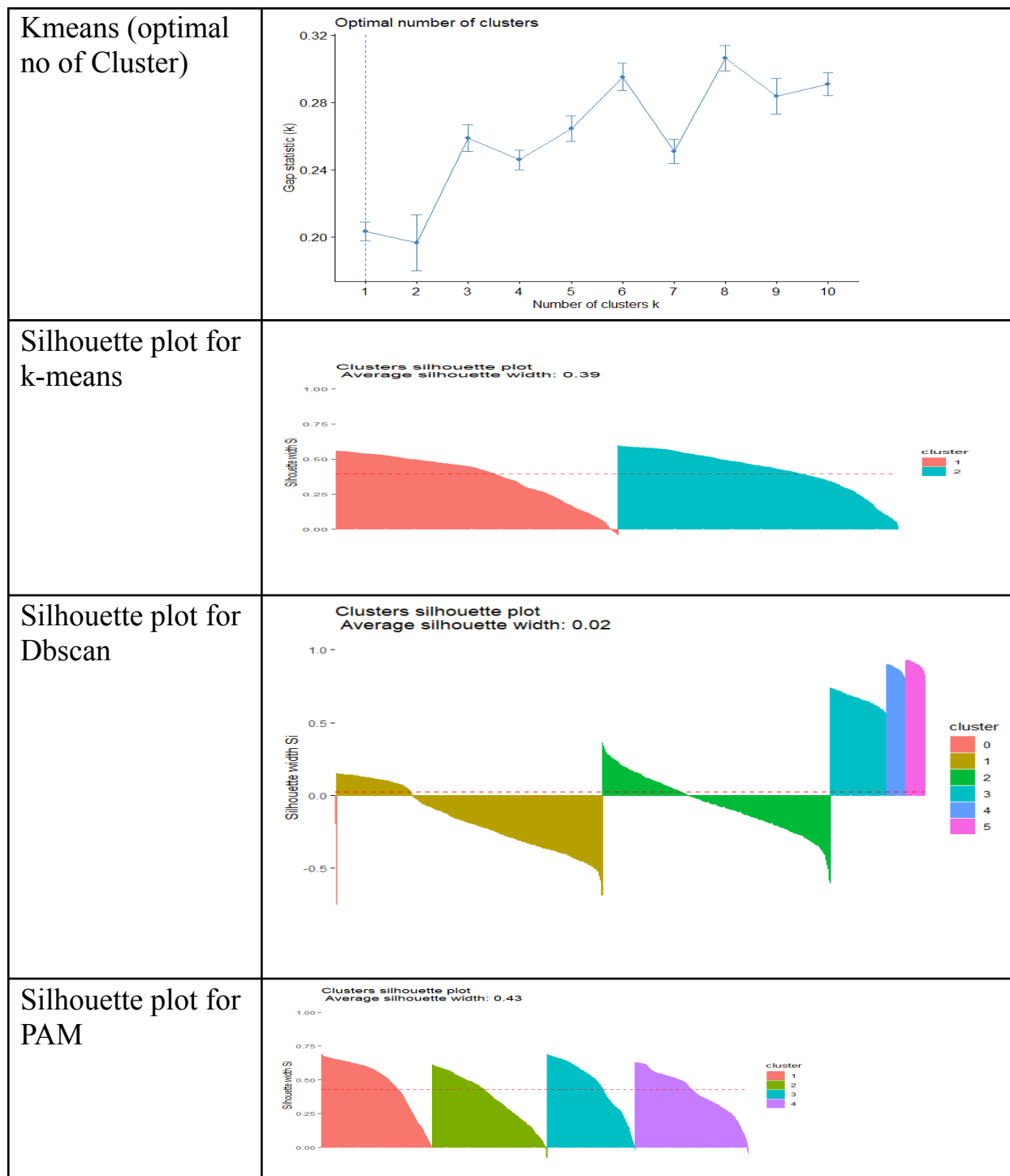


Hierarchical
al

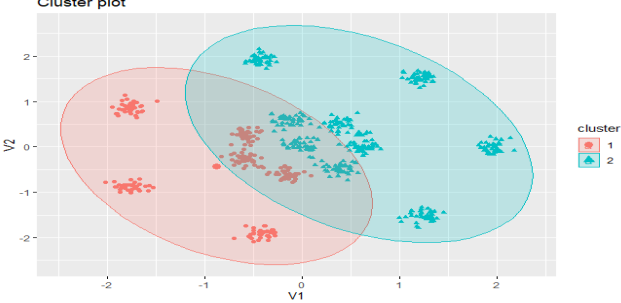
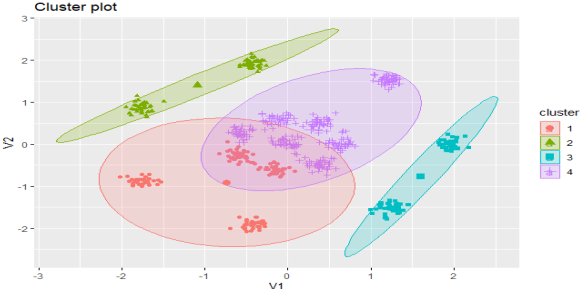
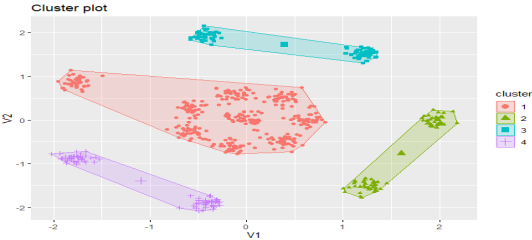




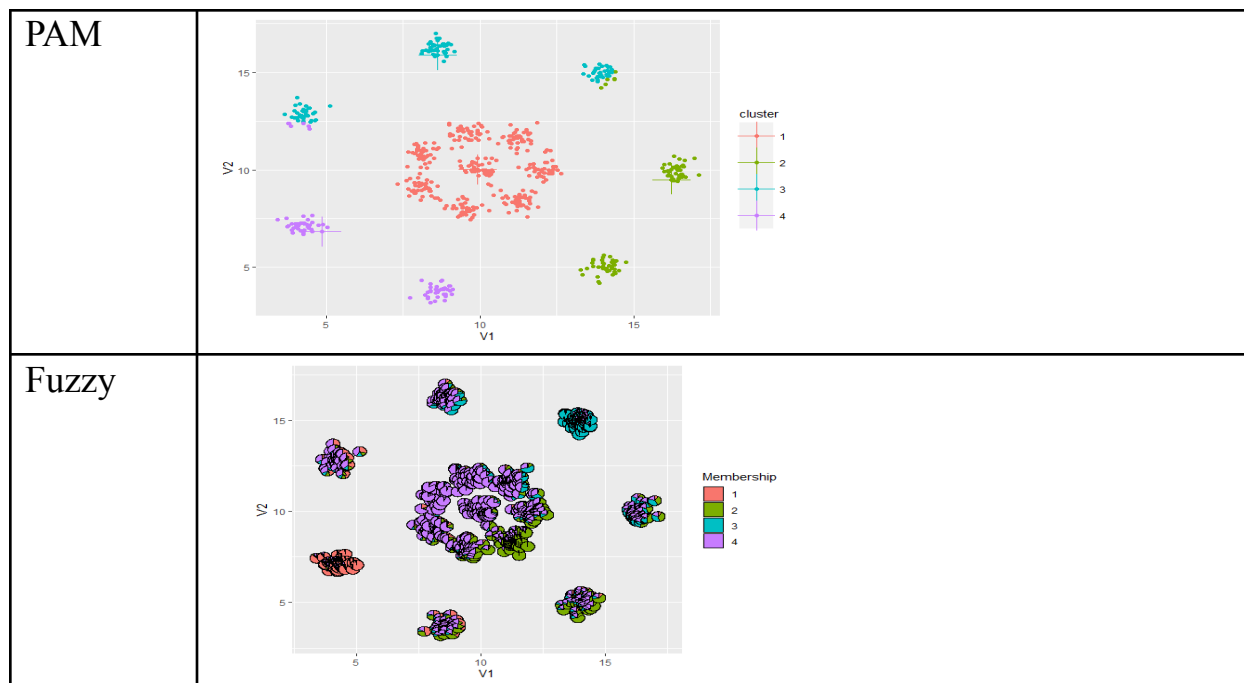
<p>Dbscan Minimum point=4</p>	
<p>Dbscan Minimum points=8</p>	
<p>PAM</p>	
<p>Fuzzy</p>	

Cluster Evaluation Shape5 :

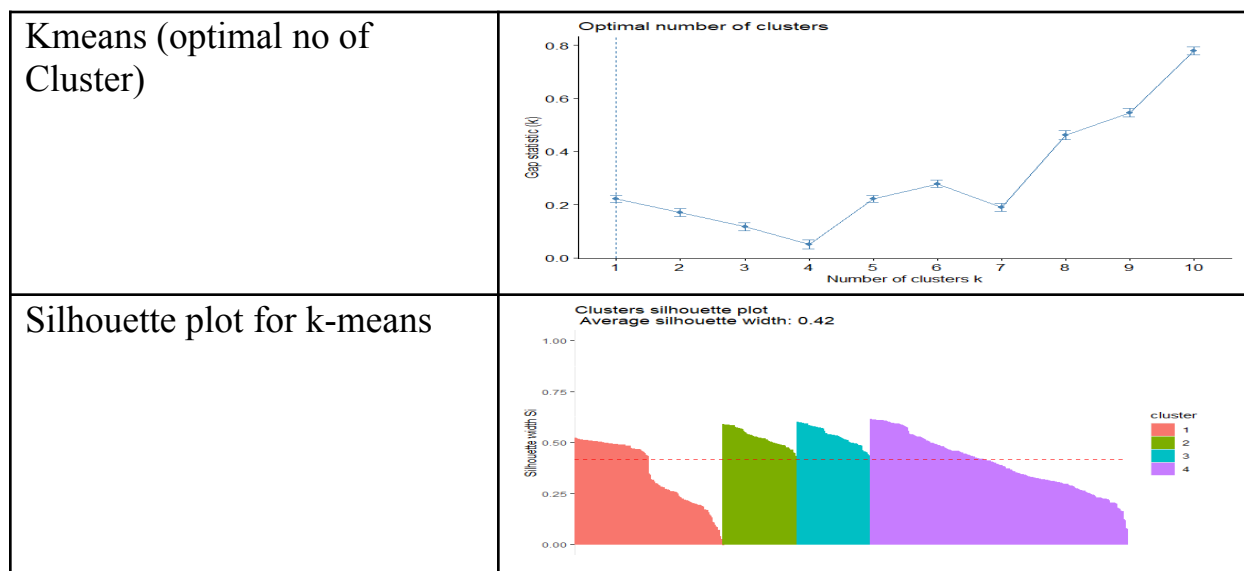


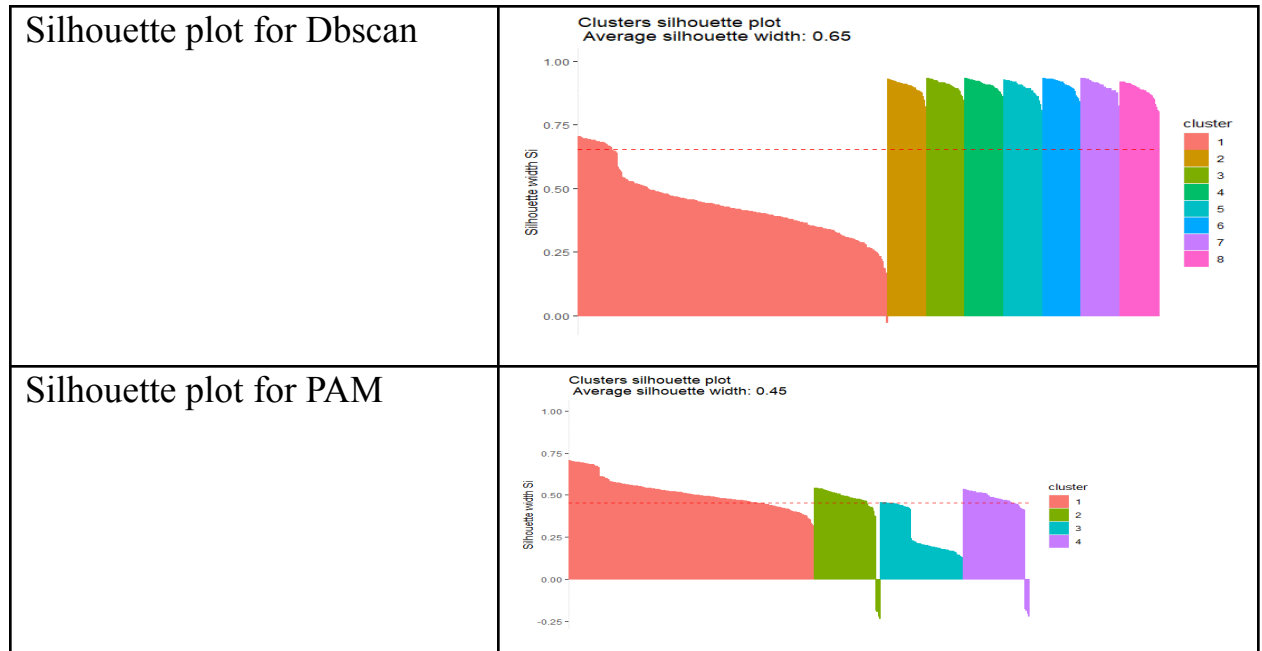
Shape 6:

<p>Kmeans(K=2)</p>	 <p>A scatter plot titled 'Cluster plot' showing two clusters of data points. The x-axis is labeled 'V1' and ranges from -2 to 2. The y-axis is labeled 'V2' and ranges from -2 to 2. Cluster 1 (red circles) is located in the lower-left region. Cluster 2 (cyan triangles) is located in the upper-right region. The clusters are separated by a decision boundary.</p>
<p>Kmeans(K=4)</p>	 <p>A scatter plot titled 'Cluster plot' showing four clusters of data points. The x-axis is labeled 'V1' and ranges from -3 to 2. The y-axis is labeled 'V2' and ranges from -2 to 3. The clusters are: Cluster 1 (red circles) in the lower-left, Cluster 2 (green triangles) in the upper-left, Cluster 3 (cyan squares) in the lower-right, and Cluster 4 (purple pluses) in the upper-middle. Each cluster is enclosed by a convex hull.</p>
<p>Hierarchical</p>	 <p>A scatter plot titled 'Cluster plot' showing four clusters of data points. The x-axis is labeled 'V1' and ranges from -2 to 2. The y-axis is labeled 'V2' and ranges from -2 to 2. The clusters are: Cluster 1 (red circles) in the lower-left, Cluster 2 (green triangles) in the upper-left, Cluster 3 (cyan squares) in the upper-right, and Cluster 4 (purple pluses) in the lower-right. The clusters are separated by a decision boundary.</p>
<p>Dbscan Minimum point=4</p>	 <p>A scatter plot showing data points clustered into 8 groups. The x-axis is labeled 'V1' and ranges from 5 to 15. The y-axis is labeled 'V2' and ranges from 5 to 15. The clusters are: Cluster 1 (red circles), Cluster 2 (yellow circles), Cluster 3 (green circles), Cluster 4 (cyan circles), Cluster 5 (blue circles), Cluster 6 (magenta circles), Cluster 7 (purple circles), and Cluster 8 (pink circles). The clusters are separated by a decision boundary.</p>
<p>Dbscan Minimum points=8</p>	 <p>A scatter plot showing data points clustered into 8 groups. The x-axis is labeled 'V1' and ranges from 5 to 15. The y-axis is labeled 'V2' and ranges from 5 to 15. The clusters are: Cluster 1 (red circles), Cluster 2 (yellow circles), Cluster 3 (green circles), Cluster 4 (cyan circles), Cluster 5 (blue circles), Cluster 6 (magenta circles), Cluster 7 (purple circles), and Cluster 8 (pink circles). The clusters are separated by a decision boundary.</p>

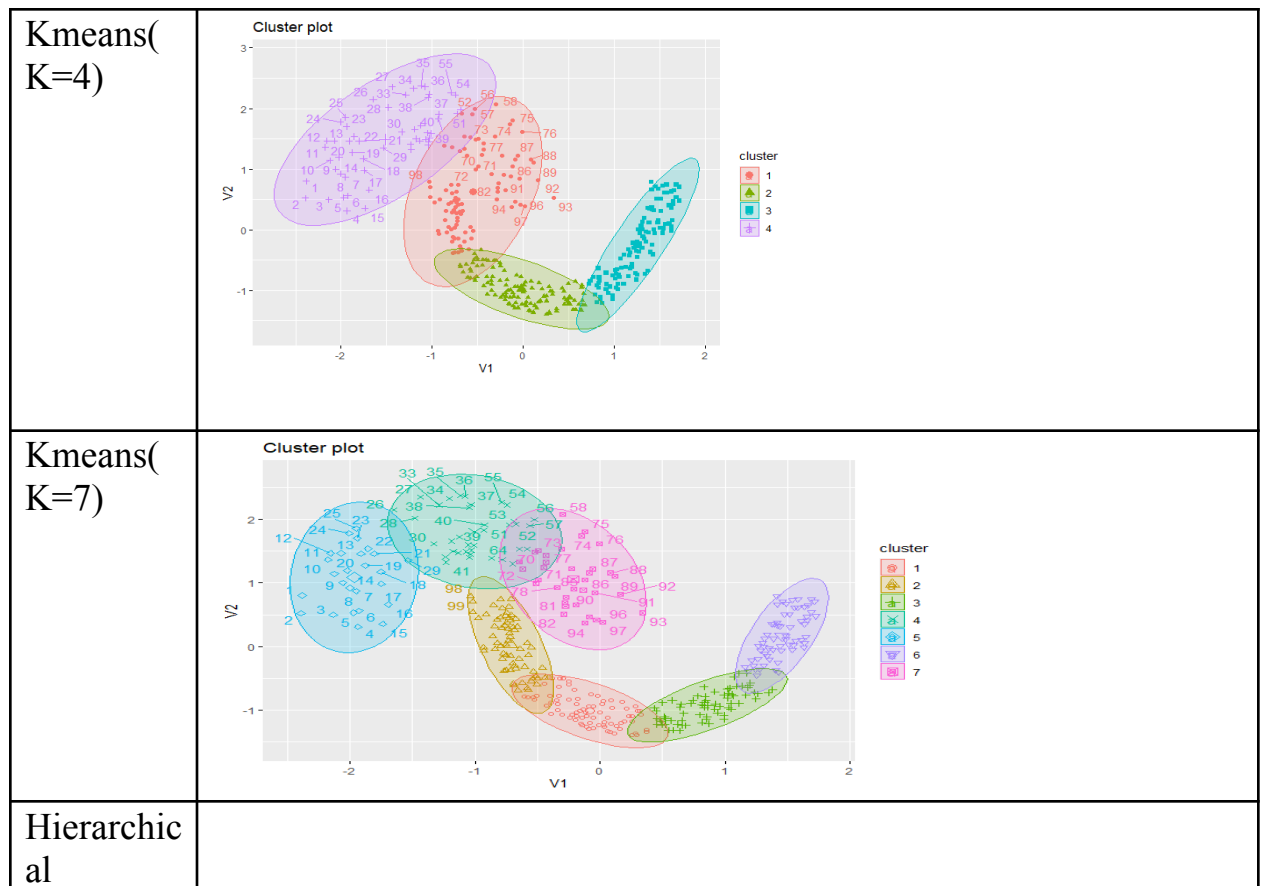


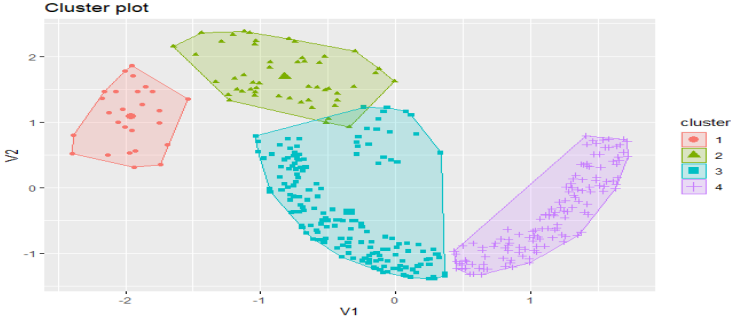

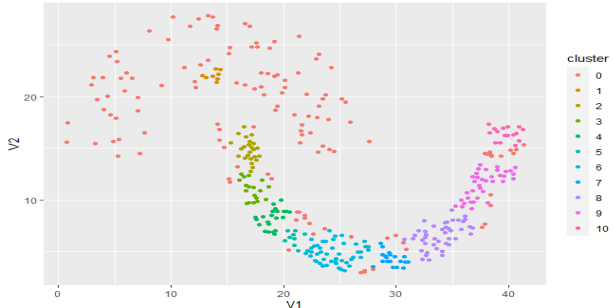

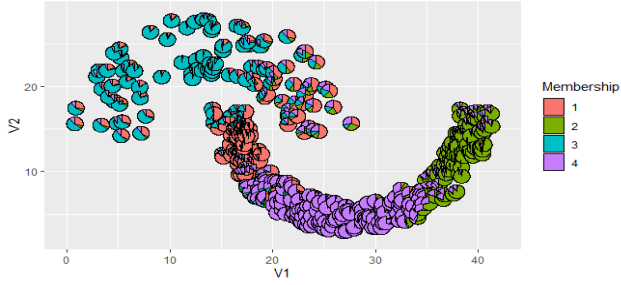
Cluster Evaluation Shape 6 :



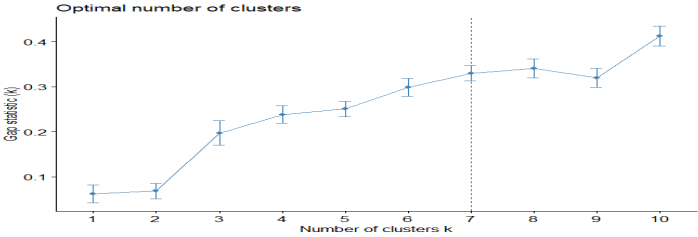
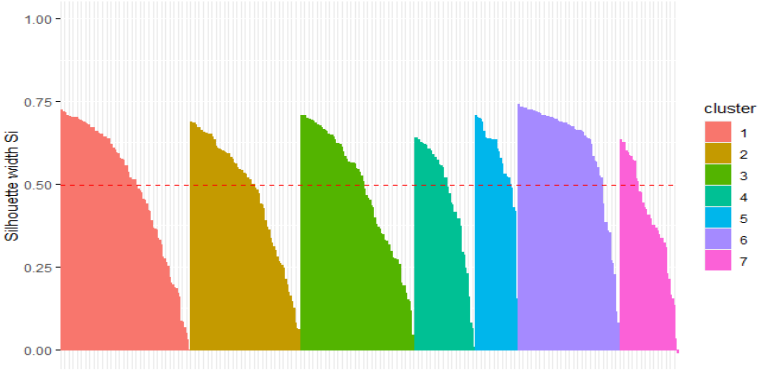
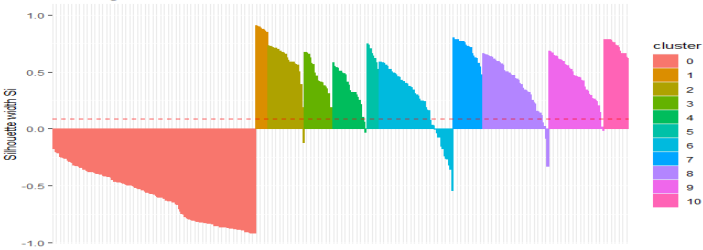
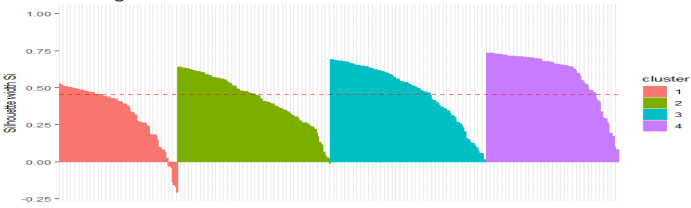


Shape 7



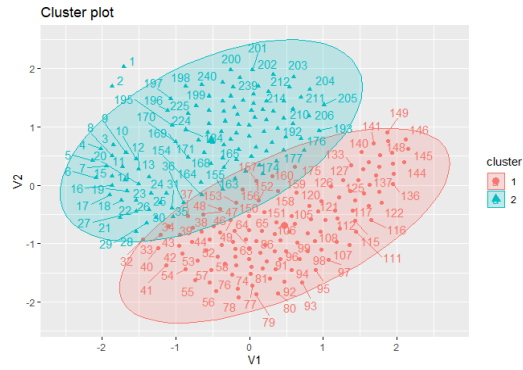
	<p>Cluster plot</p> 
Dbscan Minimum point=4	
Dbscan Minimum points=8	
PAM	
Fuzzy	

Cluster Evaluation Shape 7:

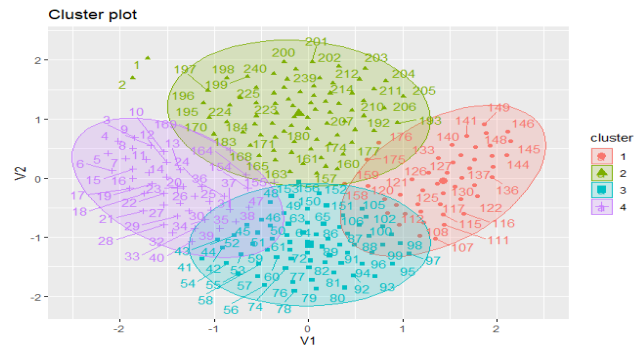
Kmeans (optimal no of Cluster)	<p>Optimal number of clusters</p>  <table border="1"><caption>Gap statistic (k) vs Number of clusters k</caption><thead><tr><th>Number of clusters k</th><th>Gap statistic (k)</th></tr></thead><tbody><tr><td>1</td><td>0.08</td></tr><tr><td>2</td><td>0.08</td></tr><tr><td>3</td><td>0.20</td></tr><tr><td>4</td><td>0.24</td></tr><tr><td>5</td><td>0.25</td></tr><tr><td>6</td><td>0.30</td></tr><tr><td>7</td><td>0.33</td></tr><tr><td>8</td><td>0.35</td></tr><tr><td>9</td><td>0.32</td></tr><tr><td>10</td><td>0.42</td></tr></tbody></table>	Number of clusters k	Gap statistic (k)	1	0.08	2	0.08	3	0.20	4	0.24	5	0.25	6	0.30	7	0.33	8	0.35	9	0.32	10	0.42		
Number of clusters k	Gap statistic (k)																								
1	0.08																								
2	0.08																								
3	0.20																								
4	0.24																								
5	0.25																								
6	0.30																								
7	0.33																								
8	0.35																								
9	0.32																								
10	0.42																								
Silhouette plot for k-means	<p>Clusters silhouette plot Average silhouette width: 0.5</p>  <table border="1"><caption>Cluster Legend for k-means</caption><thead><tr><th>cluster</th><th>color</th></tr></thead><tbody><tr><td>1</td><td>red</td></tr><tr><td>2</td><td>orange</td></tr><tr><td>3</td><td>green</td></tr><tr><td>4</td><td>teal</td></tr><tr><td>5</td><td>blue</td></tr><tr><td>6</td><td>purple</td></tr><tr><td>7</td><td>pink</td></tr></tbody></table>	cluster	color	1	red	2	orange	3	green	4	teal	5	blue	6	purple	7	pink								
cluster	color																								
1	red																								
2	orange																								
3	green																								
4	teal																								
5	blue																								
6	purple																								
7	pink																								
Silhouette plot for Dbscan	<p>Clusters silhouette plot Average silhouette width: 0.09</p>  <table border="1"><caption>Cluster Legend for DbSCAN</caption><thead><tr><th>cluster</th><th>color</th></tr></thead><tbody><tr><td>0</td><td>red</td></tr><tr><td>1</td><td>orange</td></tr><tr><td>2</td><td>yellow</td></tr><tr><td>3</td><td>green</td></tr><tr><td>4</td><td>teal</td></tr><tr><td>5</td><td>blue</td></tr><tr><td>6</td><td>purple</td></tr><tr><td>7</td><td>pink</td></tr><tr><td>8</td><td>light blue</td></tr><tr><td>9</td><td>light green</td></tr><tr><td>10</td><td>light pink</td></tr></tbody></table>	cluster	color	0	red	1	orange	2	yellow	3	green	4	teal	5	blue	6	purple	7	pink	8	light blue	9	light green	10	light pink
cluster	color																								
0	red																								
1	orange																								
2	yellow																								
3	green																								
4	teal																								
5	blue																								
6	purple																								
7	pink																								
8	light blue																								
9	light green																								
10	light pink																								
Silhouette plot for PAM	<p>Clusters silhouette plot Average silhouette width: 0.46</p>  <table border="1"><caption>Cluster Legend for PAM</caption><thead><tr><th>cluster</th><th>color</th></tr></thead><tbody><tr><td>1</td><td>red</td></tr><tr><td>2</td><td>green</td></tr><tr><td>3</td><td>teal</td></tr><tr><td>4</td><td>purple</td></tr></tbody></table>	cluster	color	1	red	2	green	3	teal	4	purple														
cluster	color																								
1	red																								
2	green																								
3	teal																								
4	purple																								

Shape 8:

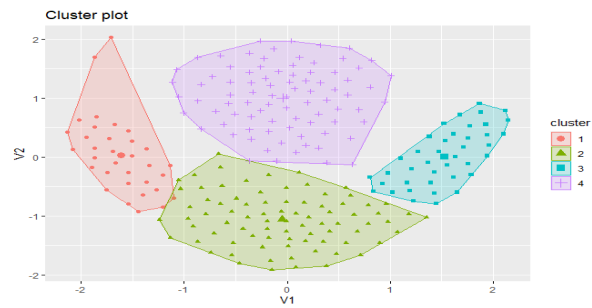
Kmeans(
K=2)



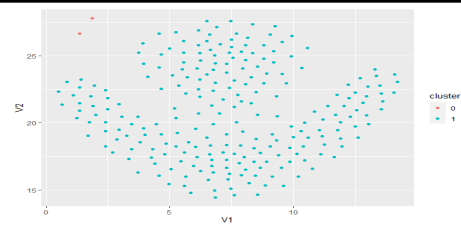
Kmeans(
K=4)

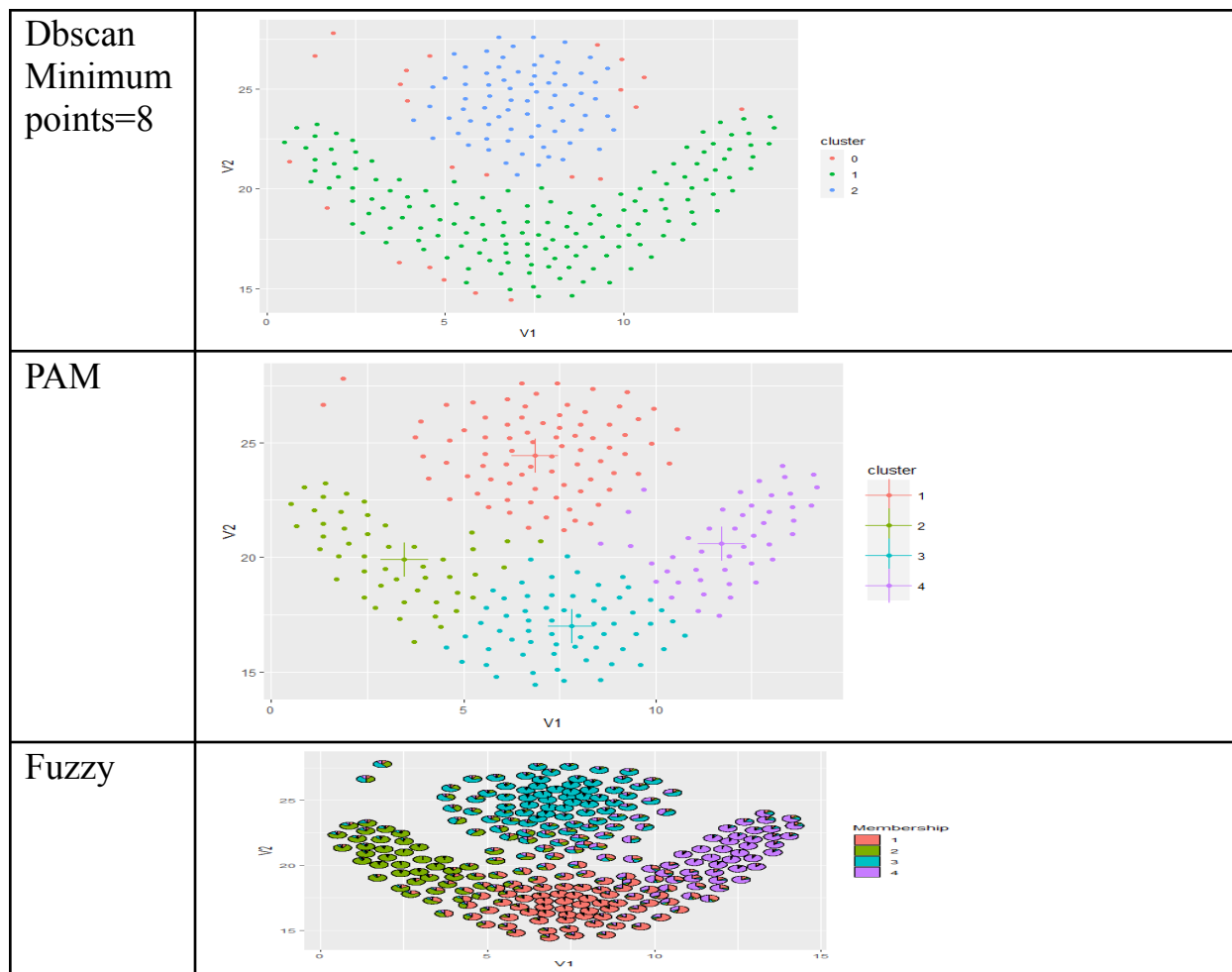


Hierarchic
al

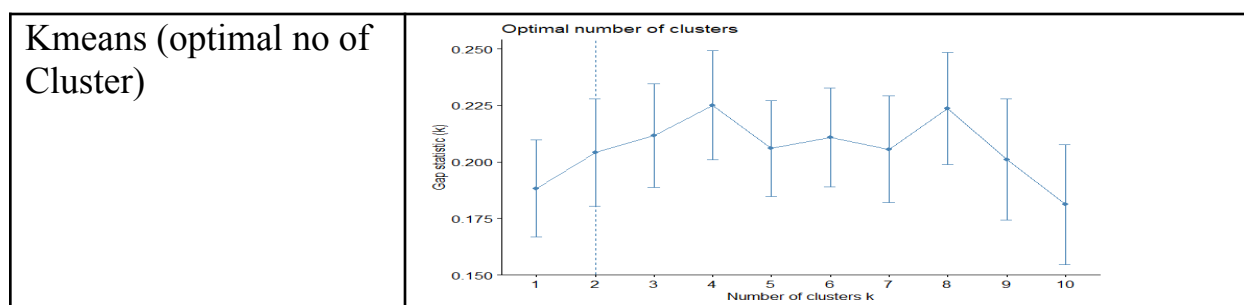


Dbscan
Minimum
point=4

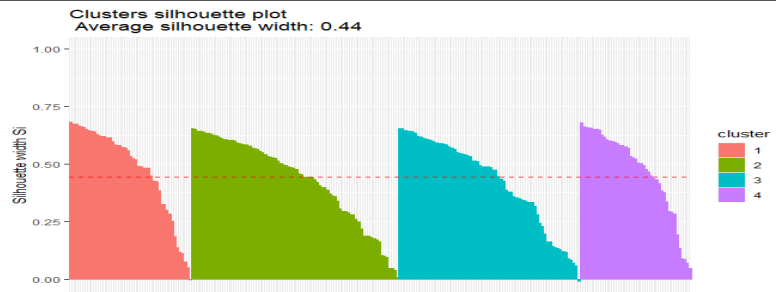




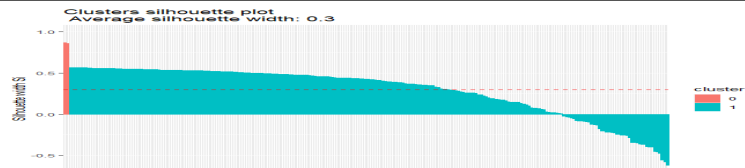
Cluster Evaluation Shape 8 :



Silhouette plot for
k-means



Silhouette plot for
Dbscan



Silhouette plot for PAM

