Part 2

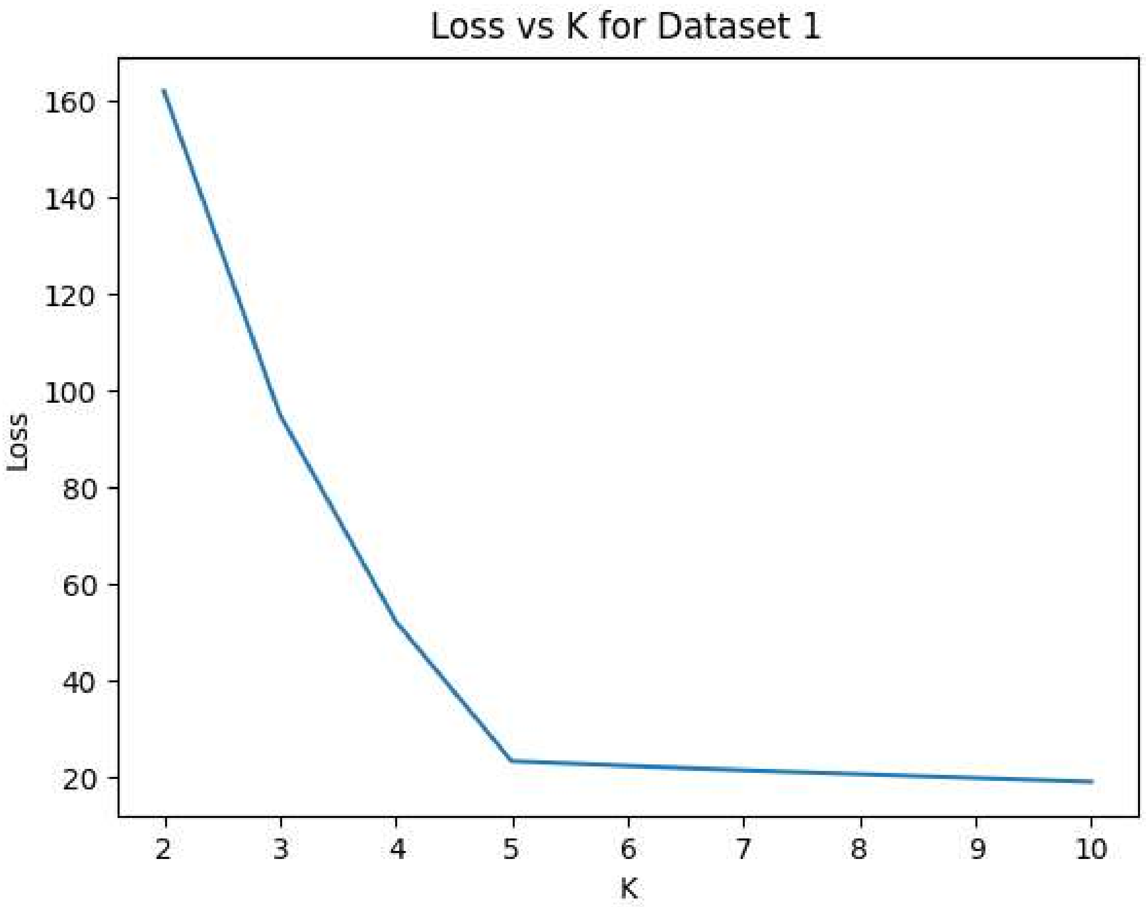
Kmeans:

Setup:

The program was run 10 times for each k-value, and then the minimum loss was taken, and this was repeated 10 times to get the average loss.

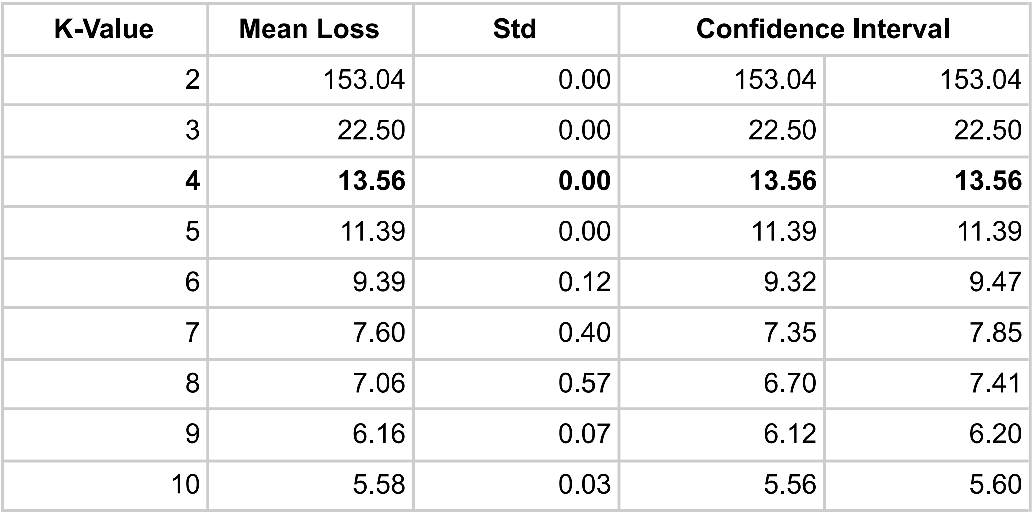
Dataset 1:

K-Value	Mean Loss	Std	Confidence Interval	
2	162.09	0.00	162.09	162.09
3	95.14	0.00	95.14	95.14
4	52.31	3.66	50.04	54.57
5	**23.37**	**0.00**	**23.37**	**23.37**
6	22.41	0.04	22.38	22.43
7	21.51	0.05	21.47	21.54
8	20.71	0.10	20.65	20.77
9	19.90	0.12	19.82	19.97
10	19.10	0.08	19.05	19.15

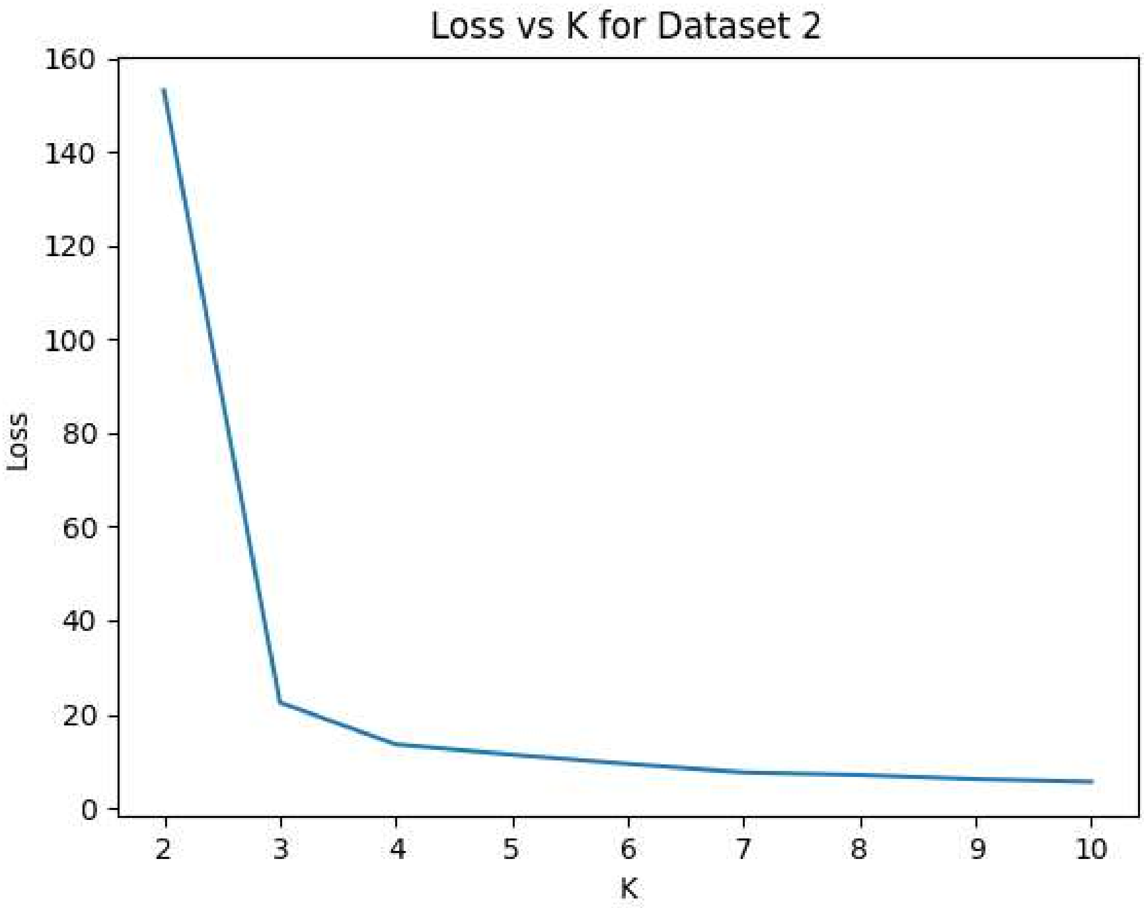


Comments: By the elbow method, we can see that the K value should be 5. From the table as well we can see that after k=5, the decrease in mean loss is very small and linear.

Dataset 2:



K-Value	Mean Loss	Std	Confidence Interval	
2	153.04	0.00	153.04	153.04
3	22.50	0.00	22.50	22.50
4	**13.56**	**0.00**	**13.56**	**13.56**
5	11.39	0.00	11.39	11.39
6	9.39	0.12	9.32	9.47
7	7.60	0.40	7.35	7.85
8	7.06	0.57	6.70	7.41
9	6.16	0.07	6.12	6.20
10	5.58	0.03	5.56	5.60



Comments: By the elbow method, we can see that the K value should be 4. From the table as well we can see that after k=4, the decrease in mean loss is very small and linear.

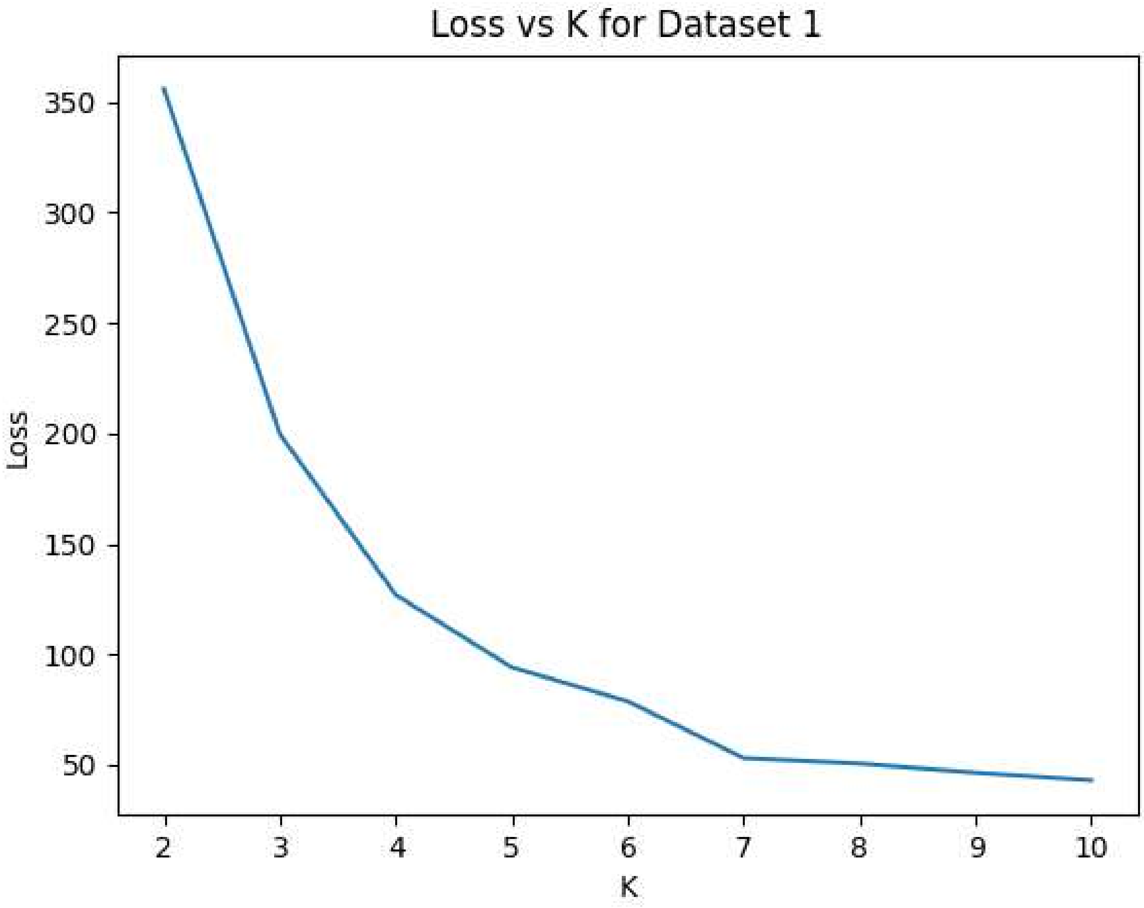
Kmedoids:

Setup:

The program was run 10 times for each k-value, and then the minimum loss was taken, and this was repeated 10 times to get the average loss.

Dataset 1:

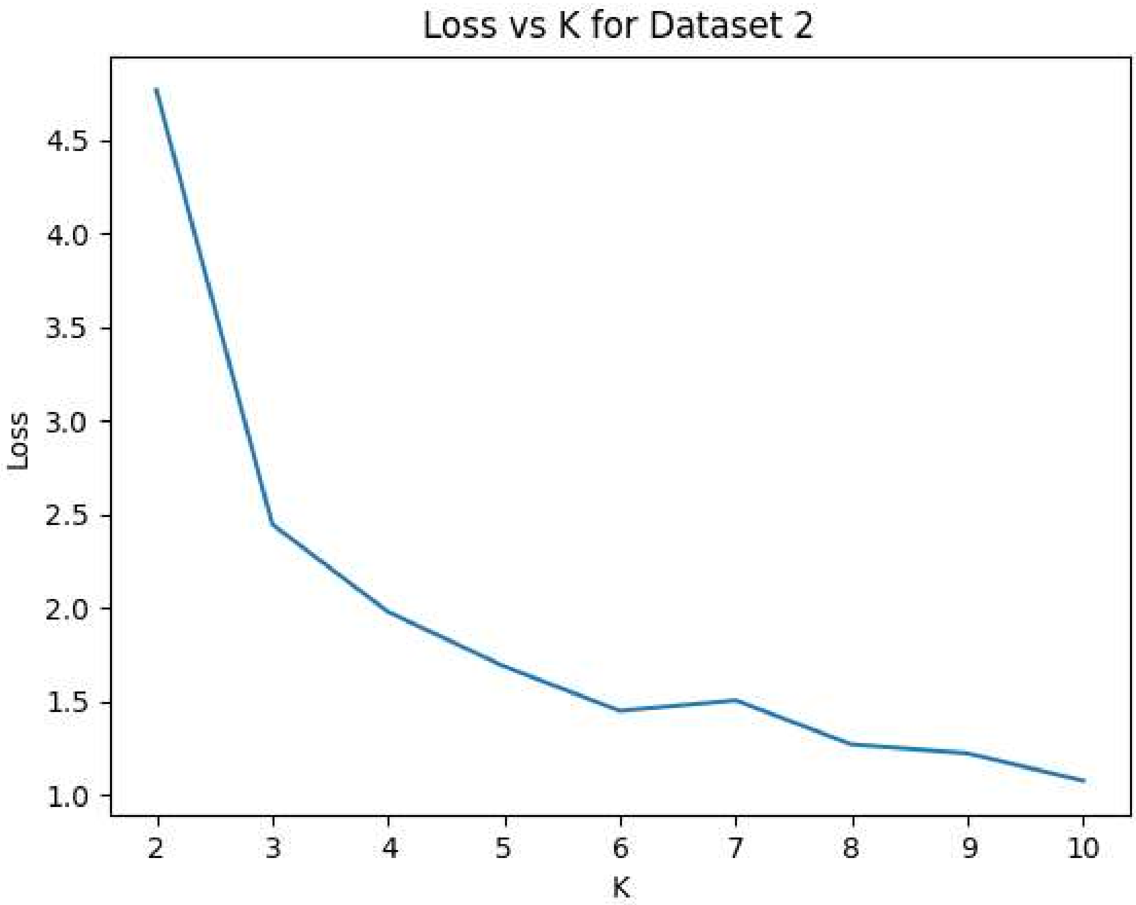
K-Value	Mean Loss	Std	Confidence Interval	
2	355.86	10.22	349.52	362.20
3	199.75	23.23	185.35	214.15
4	126.97	19.22	115.06	138.88
5	94.23	15.85	84.40	104.05
6	78.86	18.21	67.58	90.15
7	53.17	3.97	50.71	55.63
8	50.67	5.38	47.33	54.00
9	46.56	2.55	44.98	48.15
10	43.10	2.53	41.54	44.67



Comments: By the elbow method, we can see that the K value should be 5. From the table as well we can see that after k=5, the decrease in mean loss is very small and linear.

Dataset 2:

K-Value	Mean Loss	Std	Confidence Interval	
2	4.75	0.73	4.30	5.20
3	2.67	0.75	2.21	3.13
4	2.00	0.22	1.86	2.14
5	1.62	0.10	1.56	1.68
6	1.46	0.15	1.37	1.55
7	1.36	0.10	1.30	1.43
8	1.27	0.12	1.20	1.35
9	1.15	0.08	1.10	1.20
10	1.14	0.05	1.11	1.17



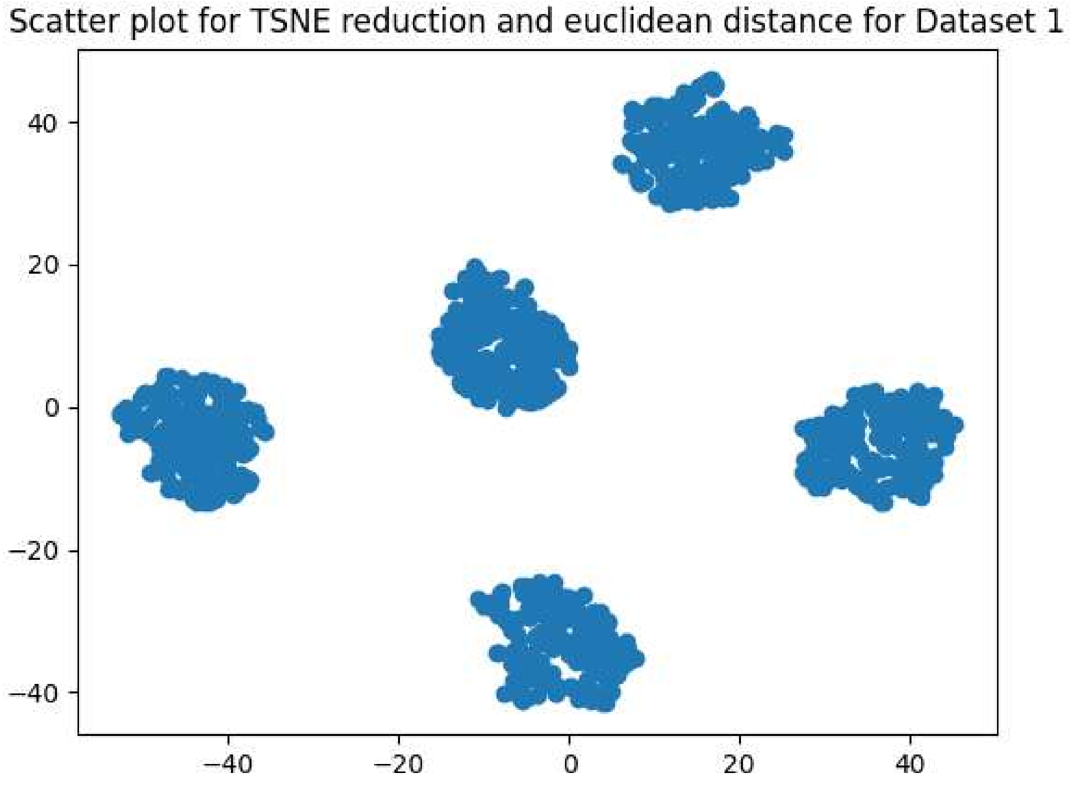
Comments: By the elbow method, we can see that the K value should be 4. From the table as well we can see that after k=4, the decrease in mean loss is very small and linear.

Dimensionality Reduction:

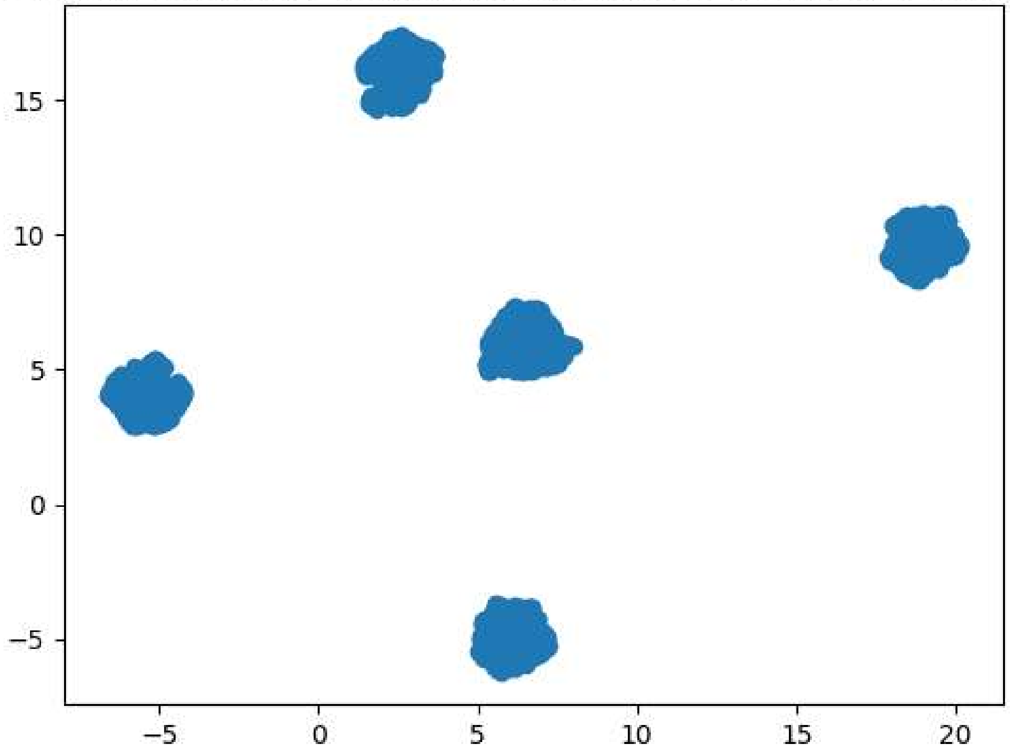
Setup:

Dimensionality reduction is applied for both datasets with TSNE and UMAP and both distance metrics, euclidean and cosine.

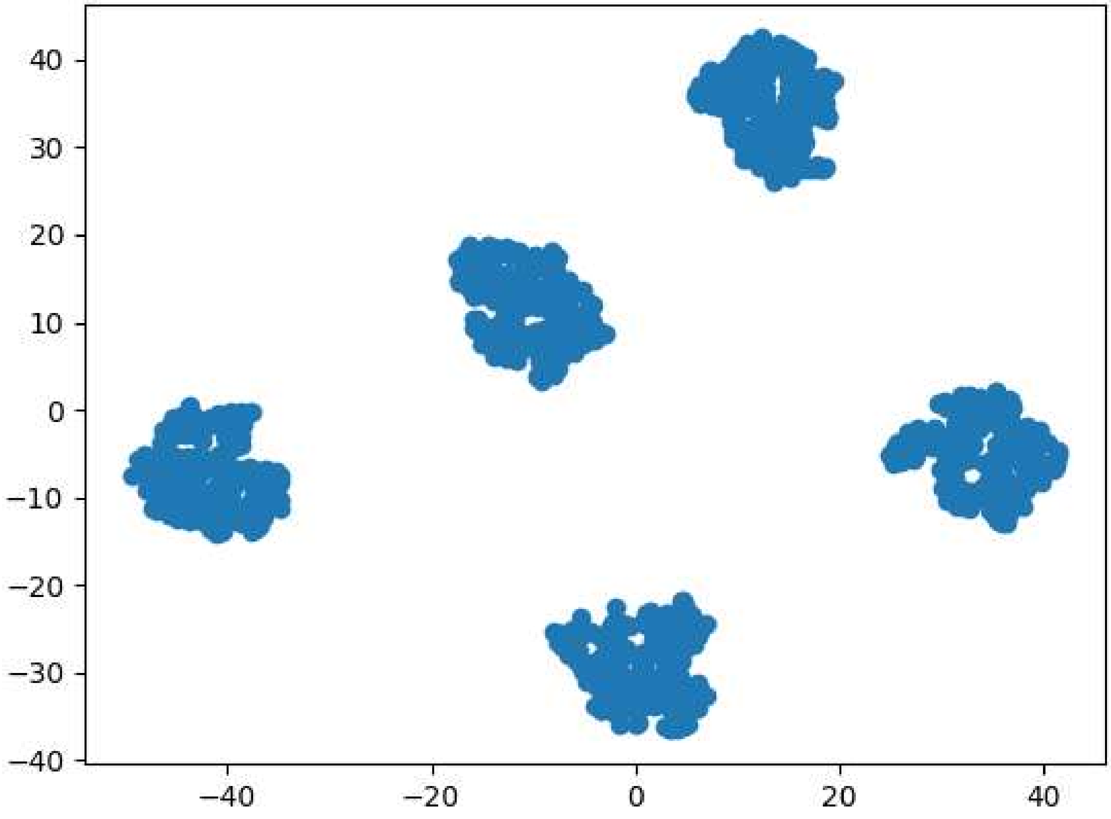
Dataset 1:

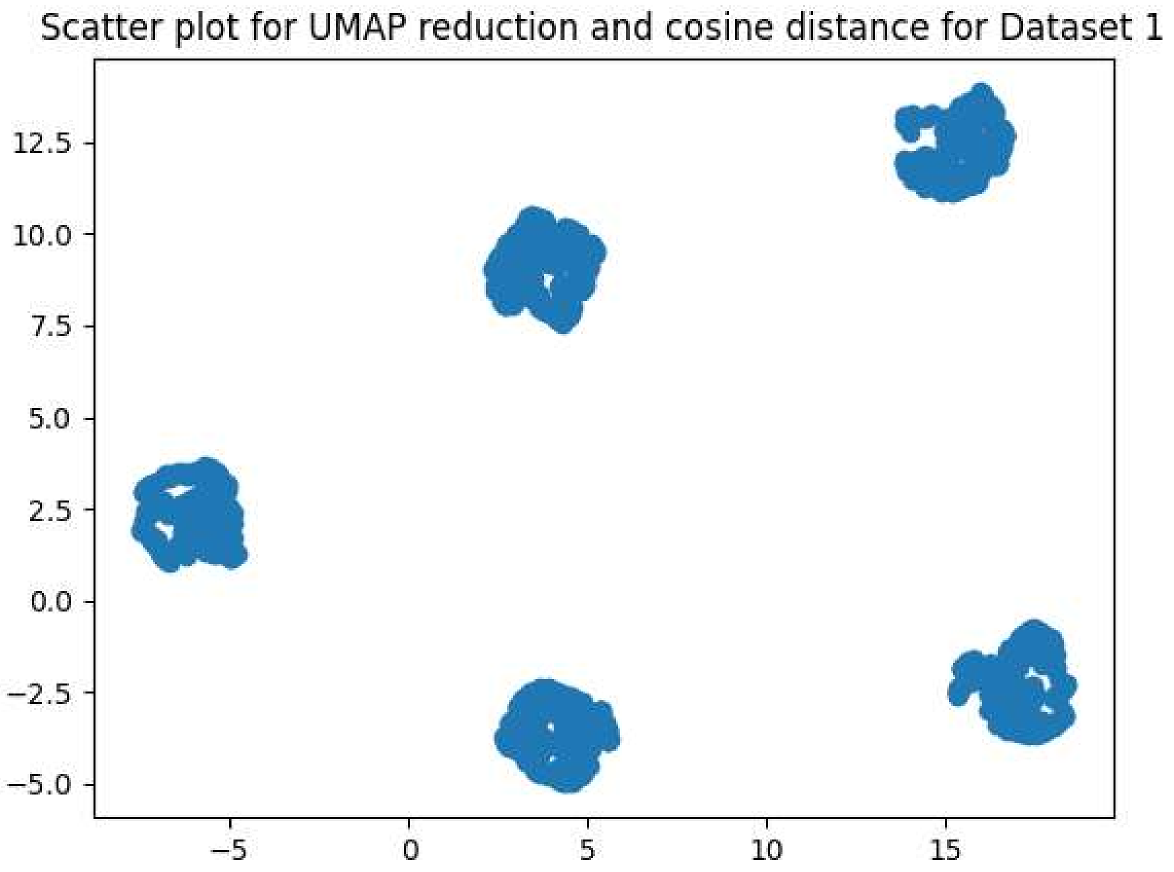


Scatter plot for UMAP reduction and euclidean distance for Dataset 1



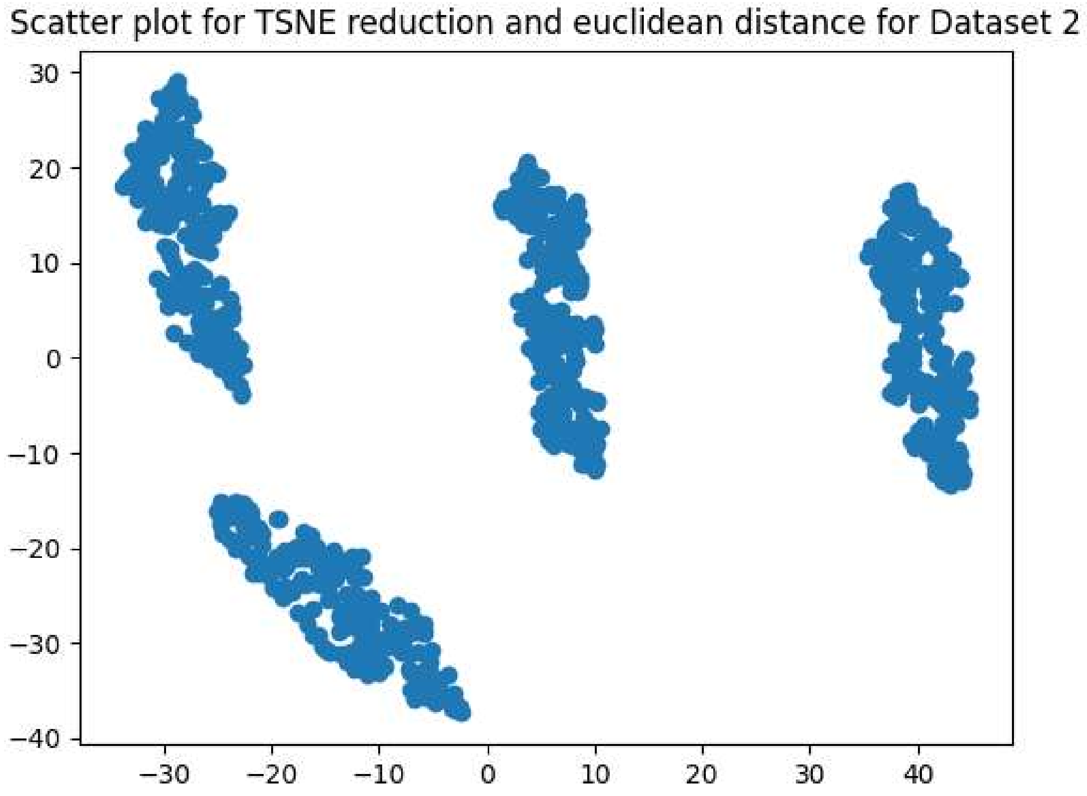
Scatter plot for TSNE reduction and cosine distance for Dataset 1

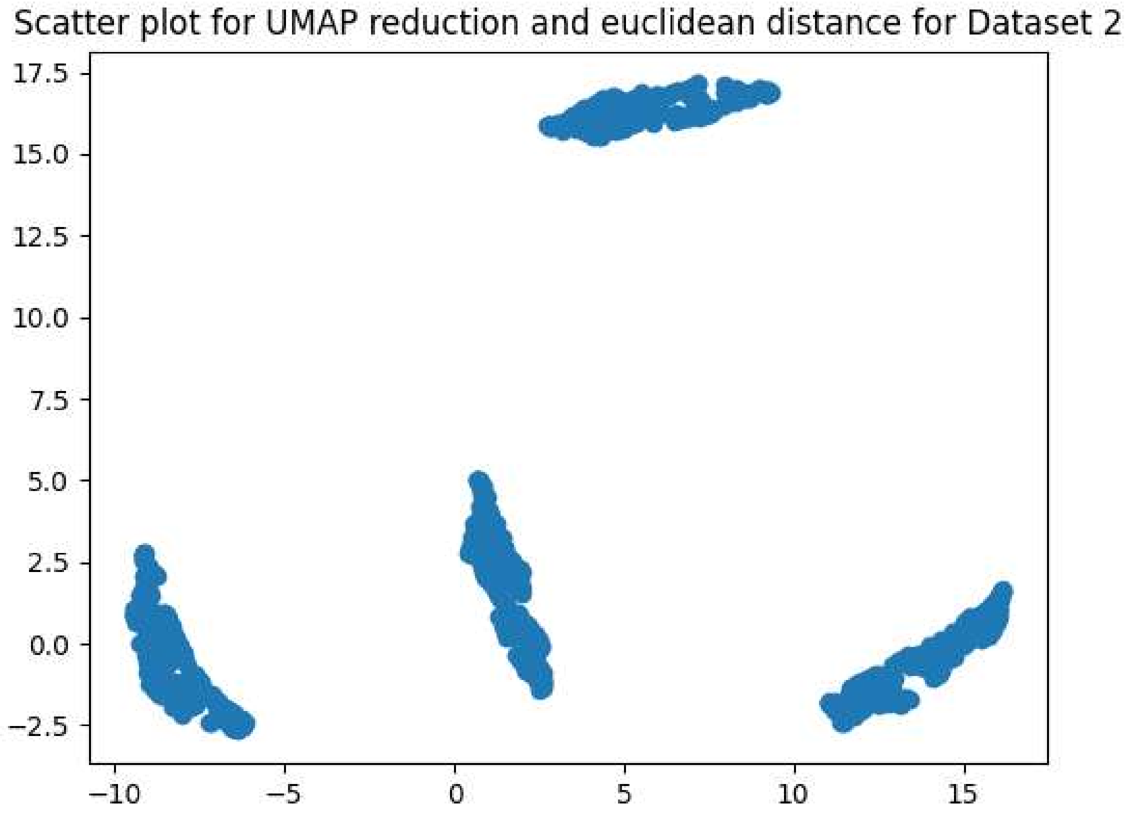


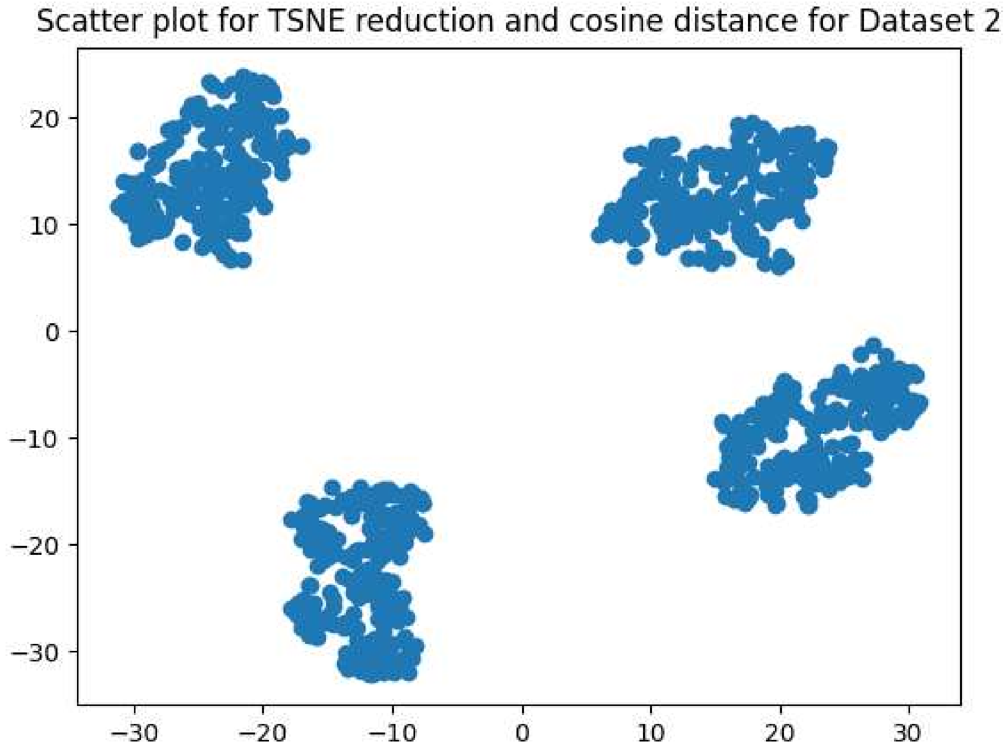


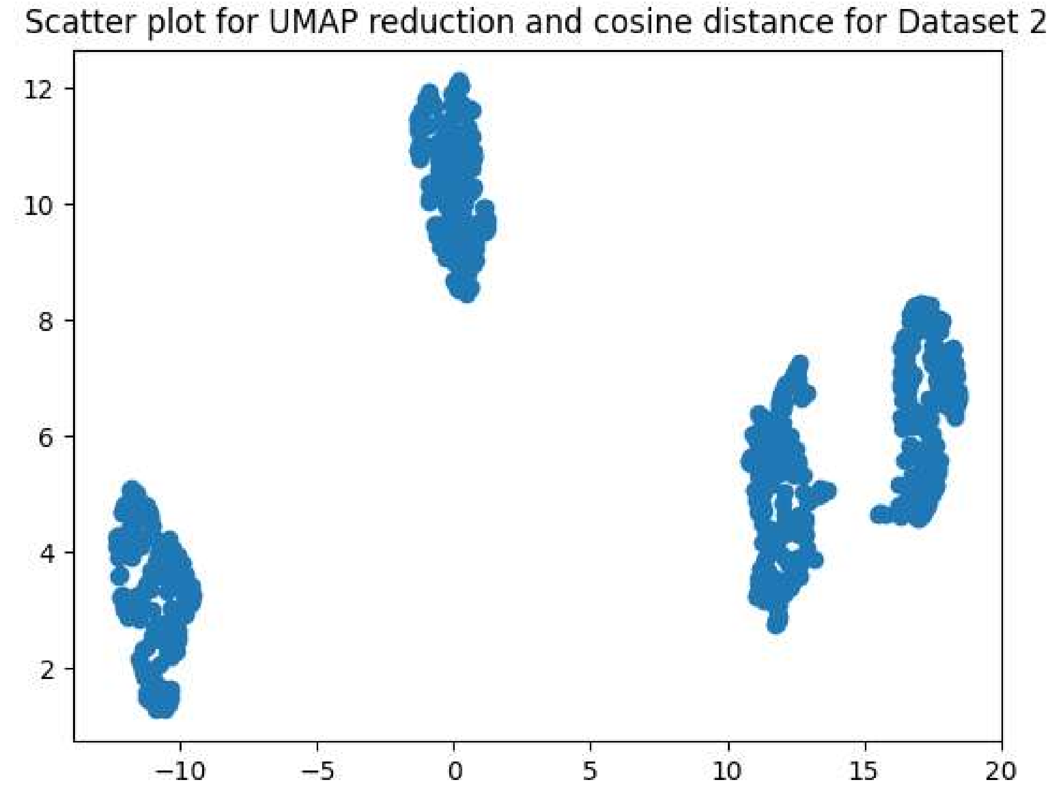
Comments: All plots show that the clusters should be 5. In my opinion, the best one is TSNE with cosine distance as the points are not too much compressed, which shows that they have enough data and the clusters are compact.

Dataset 2:


Scatter plot for TSNE reduction and euclidean distance for Dataset 2
30
20
10
0
−10
−20
−30
−40
−30
−20
−10
0
10
20
30
40


Scatter plot for UMAP reduction and euclidean distance for Dataset 2
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−10
−5
0
5
10
15





Comments: All plots show that the clusters should be 4. In my opinion, the best one is TSNE with Euclidean distance as the points are not too much compressed, which shows that they have enough data and the clusters are compact.

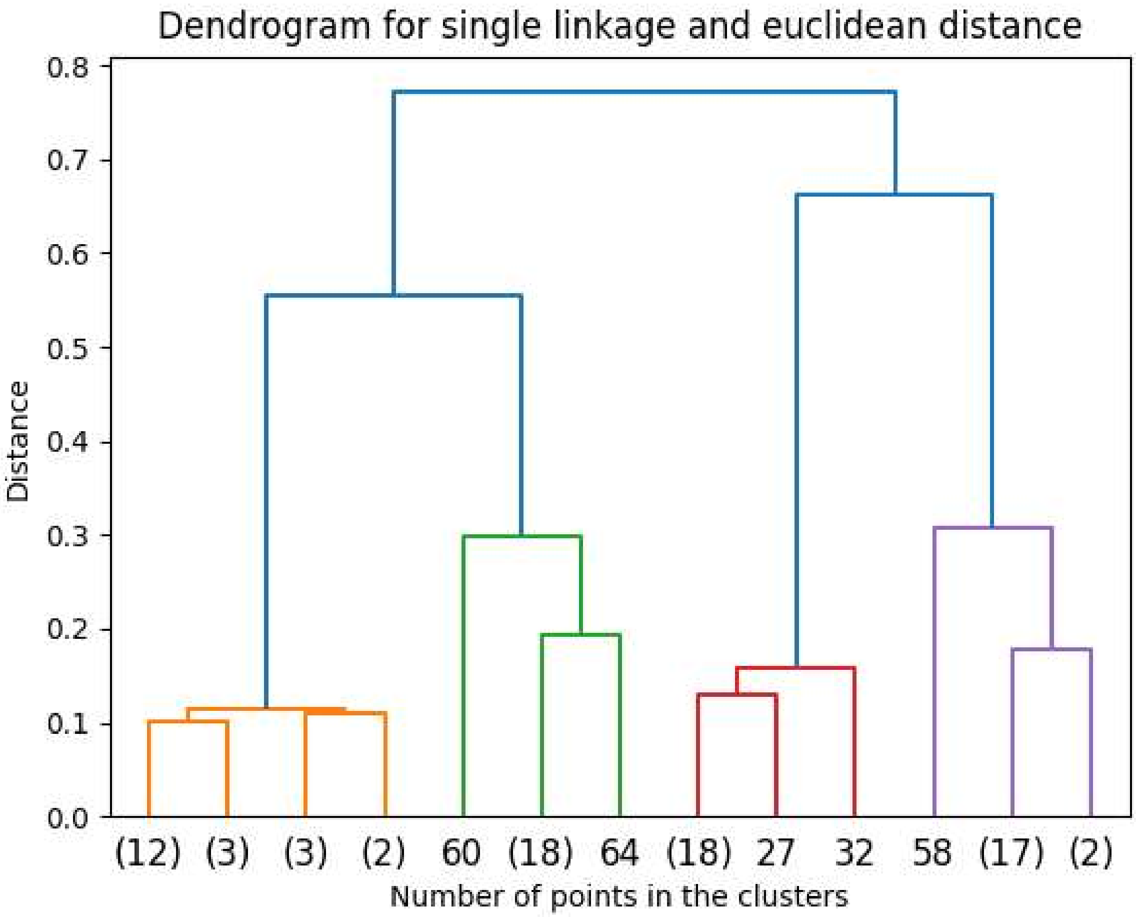
Worst Case Analysis:

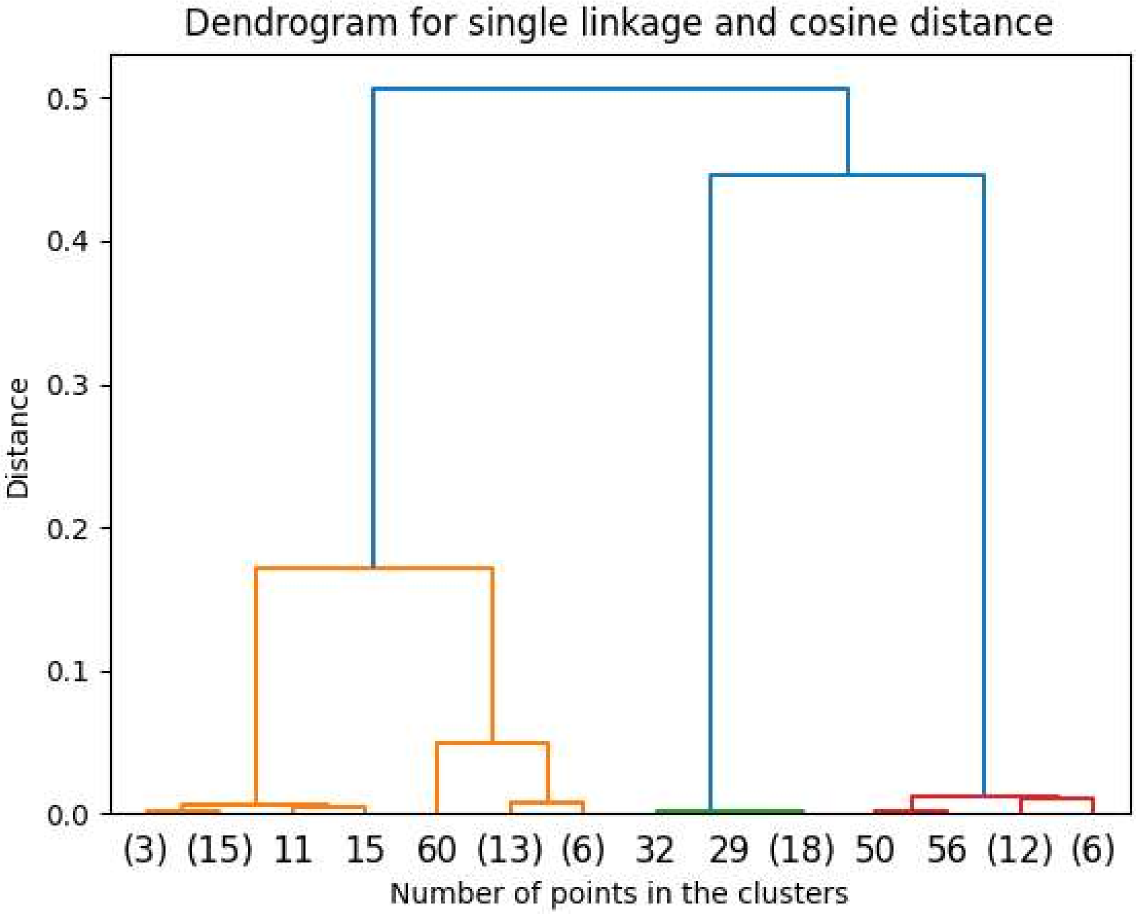
Worst-case running time analysis with respect to the number of data points (N), data sample vector dimension (d), cluster number (K), and the number of iterations (I).

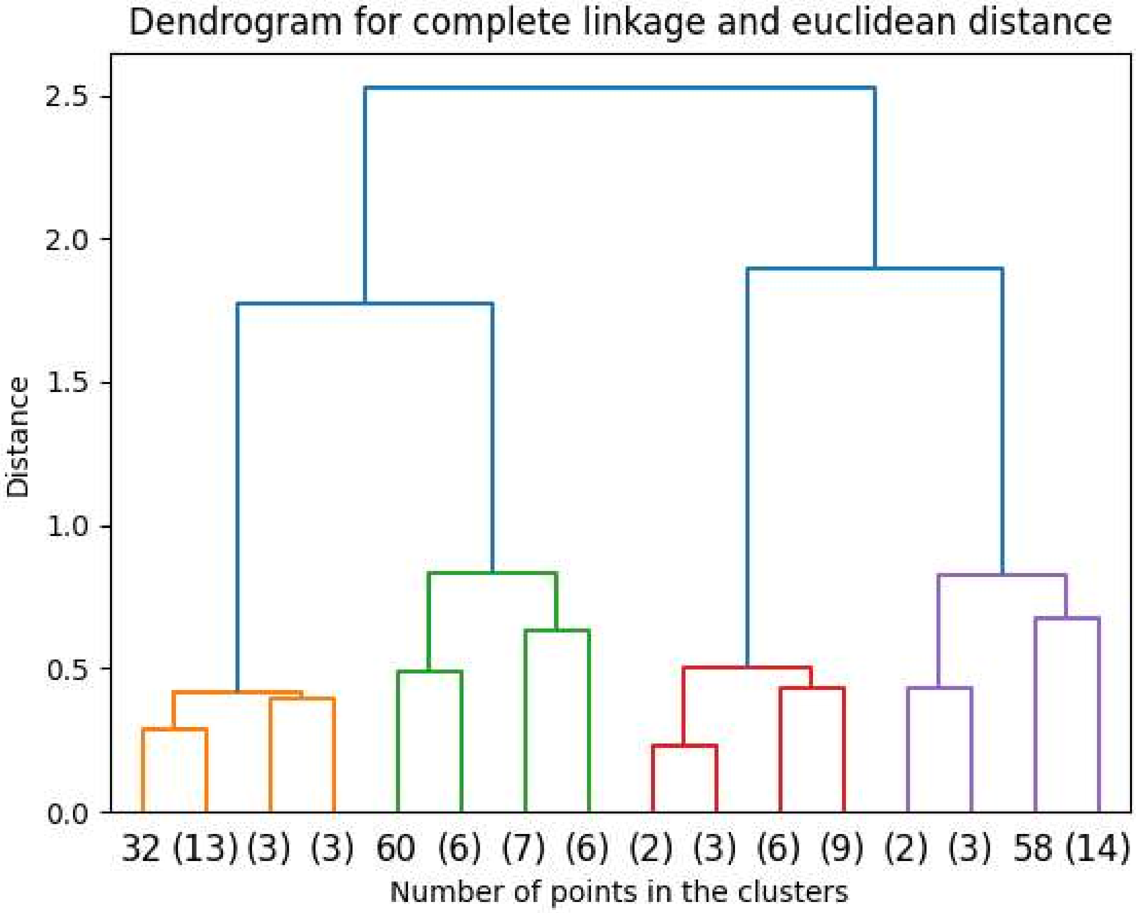
For expression 1 in assignment N, data points will be iteration, and setting the clusters according to centers will take O(K). The distance will be found for each cluster with each data point, so it will become O(NKd). If we do this for iteration I, the final worst time will be **O(IKND).**

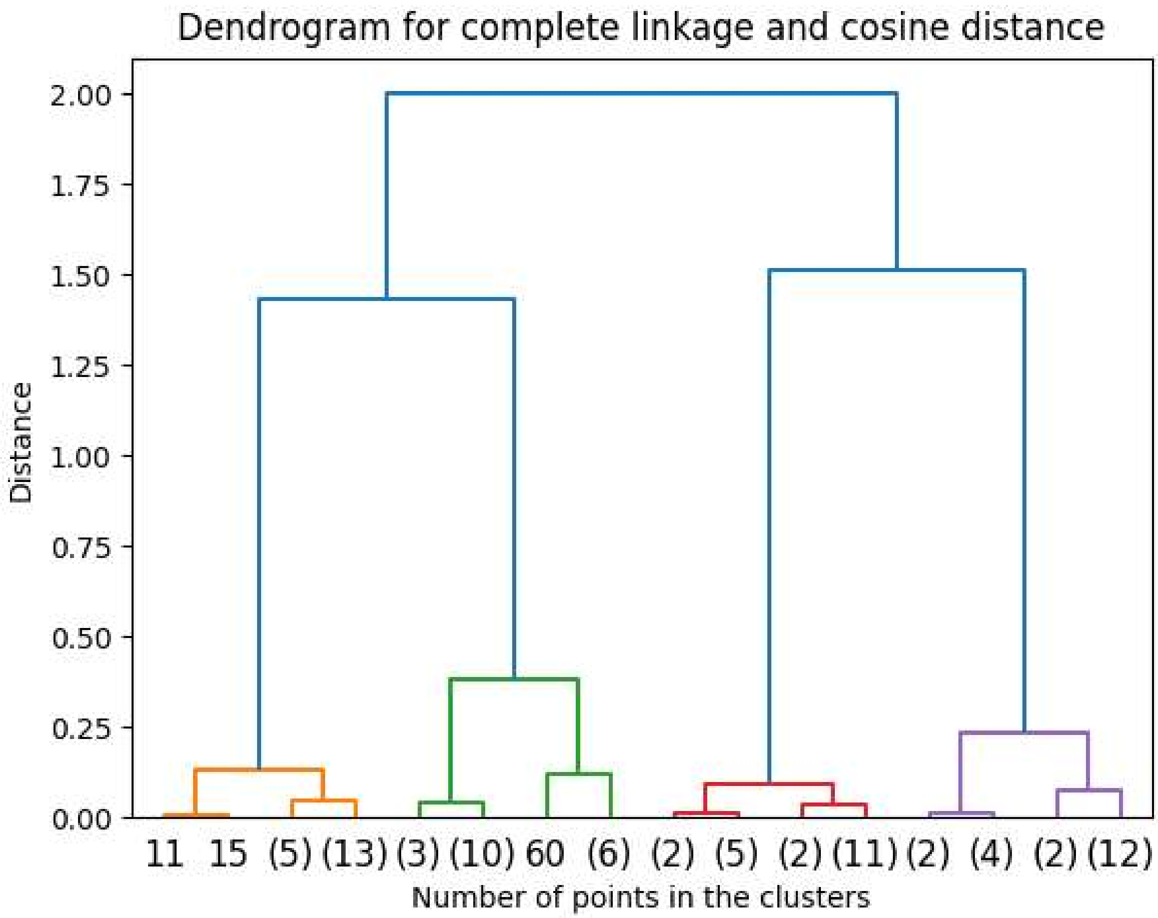
<u>Part 3</u>

HAC:


Dendrogram for single linkage and euclidean distance
Distance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
(12)
(3)
(3)
(2)
60
(18)
64
(18)
27
32
58
(17)
(2)
Number of points in the clusters

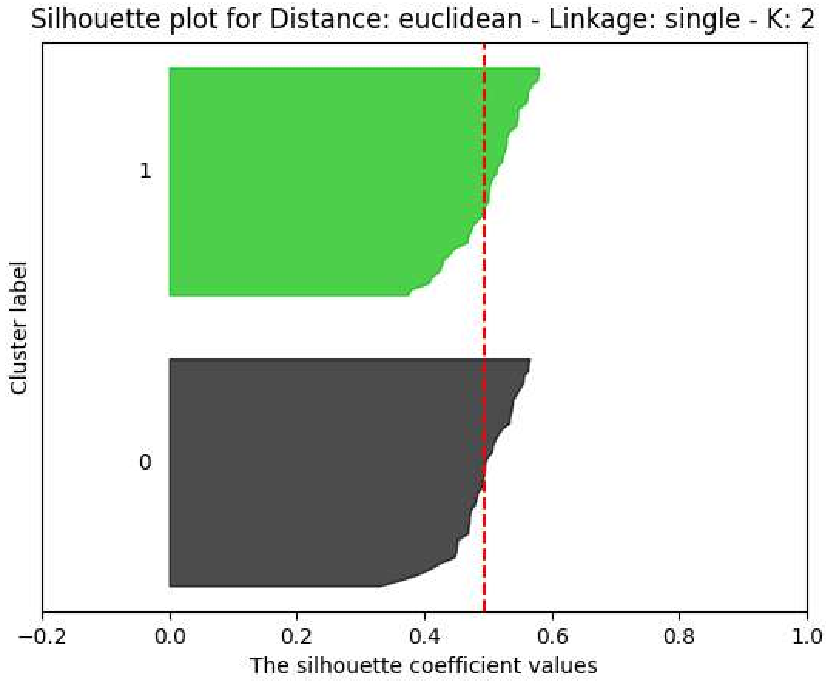

Dendrogram for single linkage and cosine distance
Distance
0.0
0.1
0.2
0.3
0.4
0.5
(3) (15) 11 15 60 (13) (6) 32 29 (18) 50 56 (12) (6)
Number of points in the clusters


Dendrogram for complete linkage and euclidean distance
Distance
0.0
0.5
1.0
1.5
2.0
2.5
32 (13) (3) (3) 60 (6) (7) (6) (2) (3) (6) (9) (2) (3) 58 (14)
Number of points in the clusters

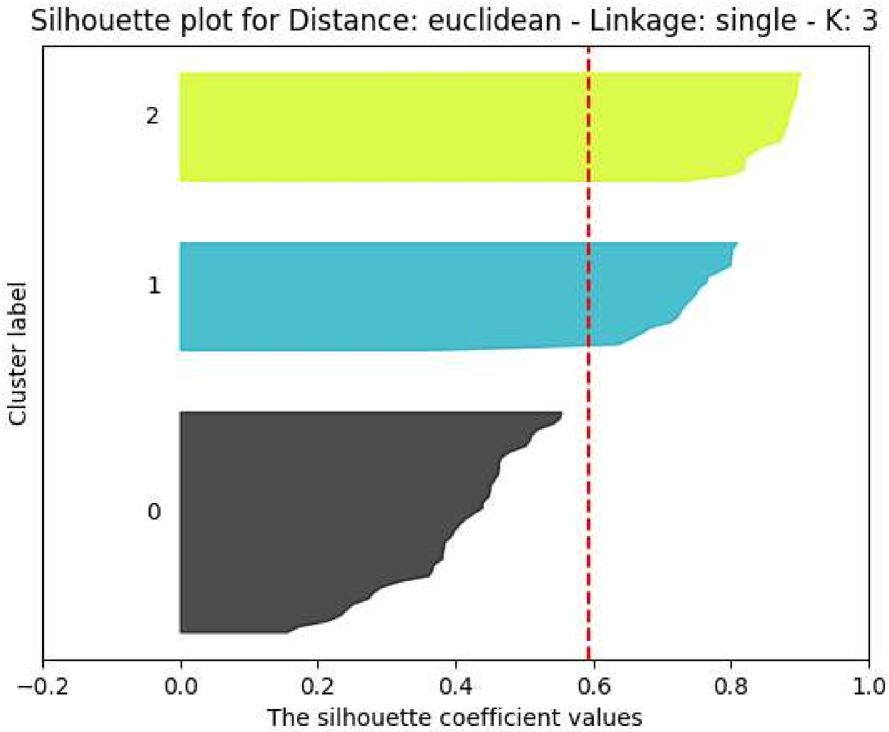


Comments: Every HAC shows that there should be 4 clusters

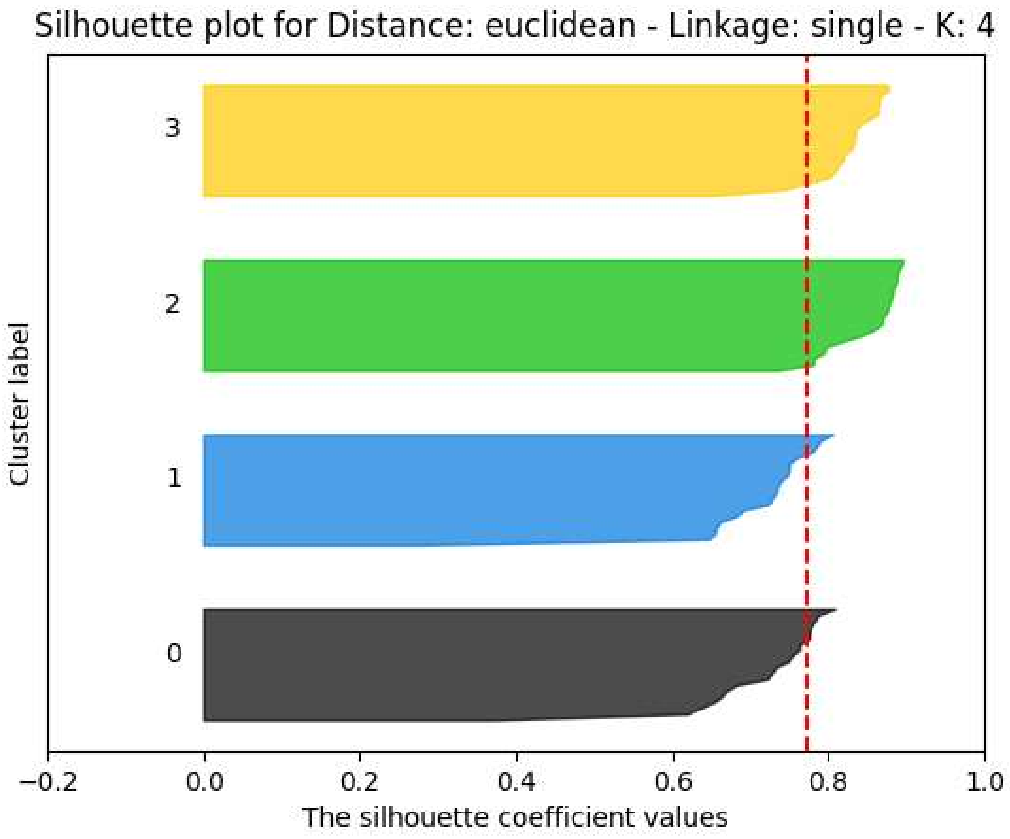
Silhouette Plot:



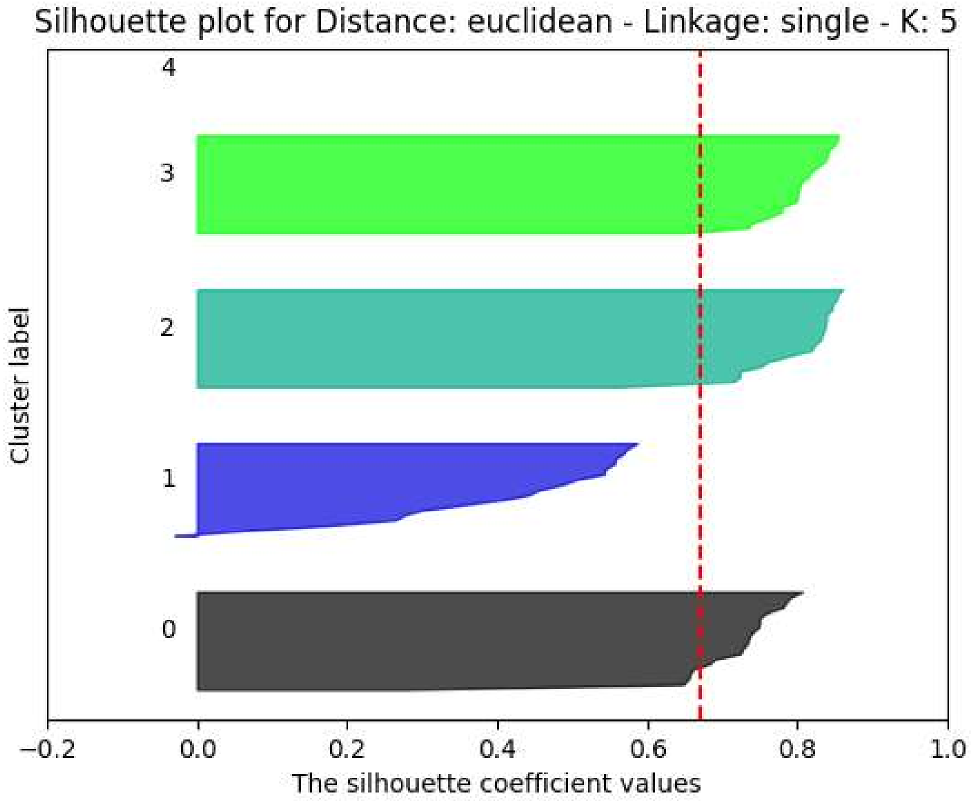
Comments: This shows good clustering.



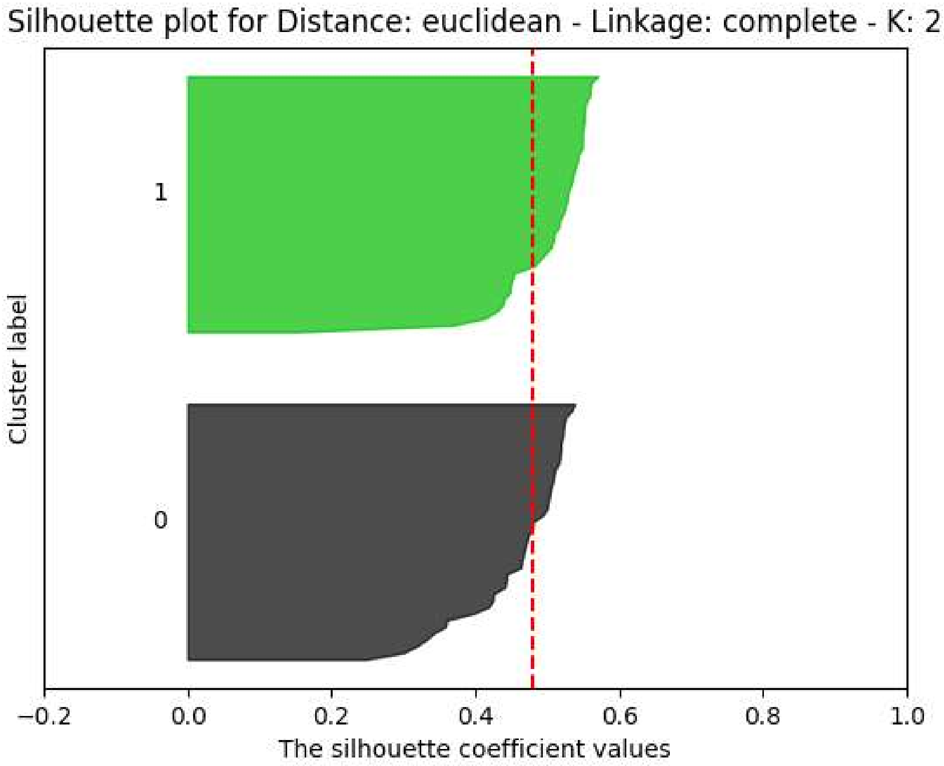
Comments: This shows good clustering.



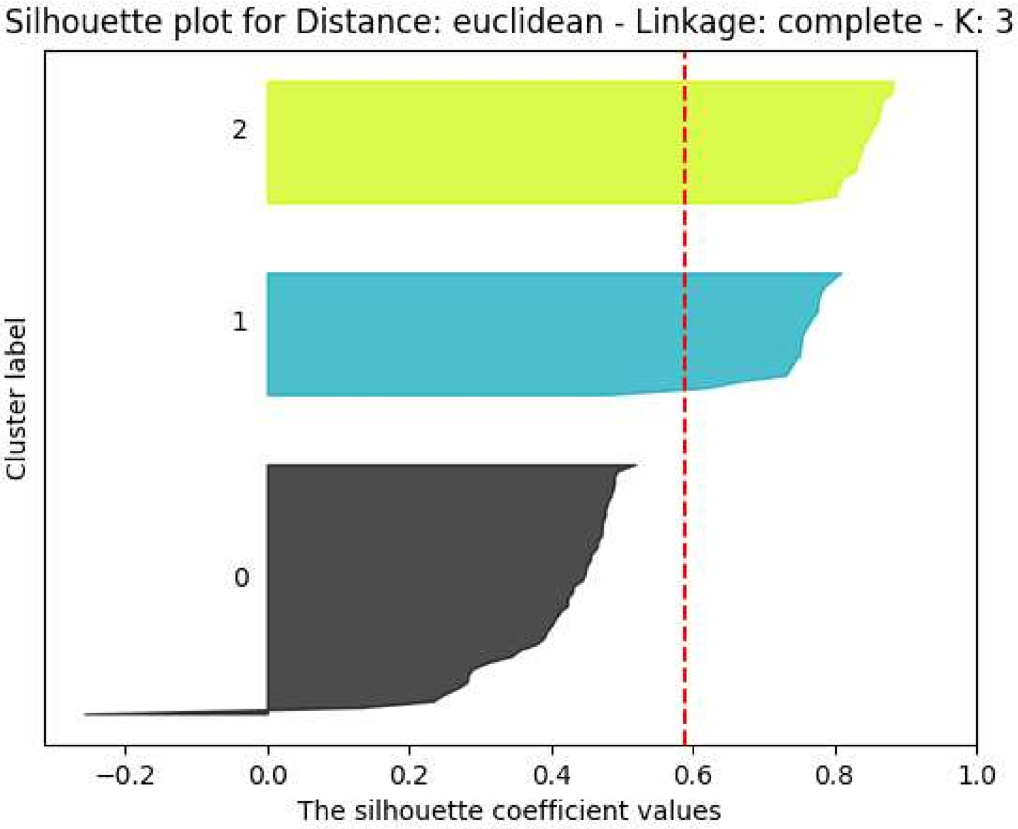
Comments: This shows good clustering.



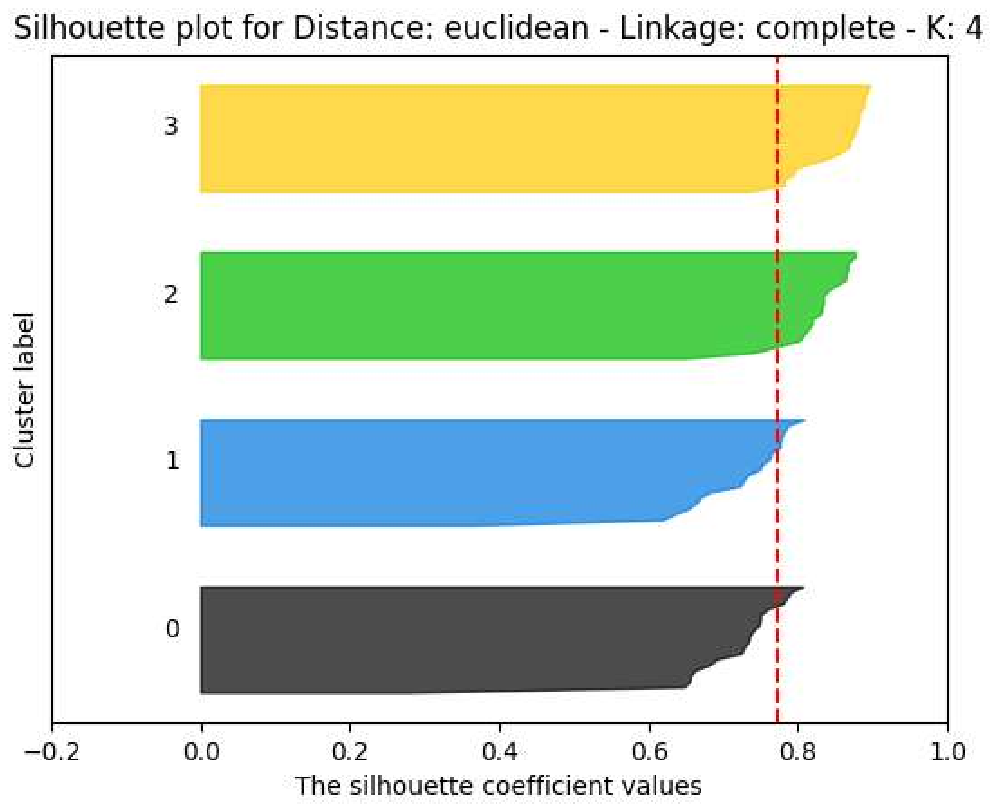
Comments: Label 1 has negative values which shows it has misclassification.



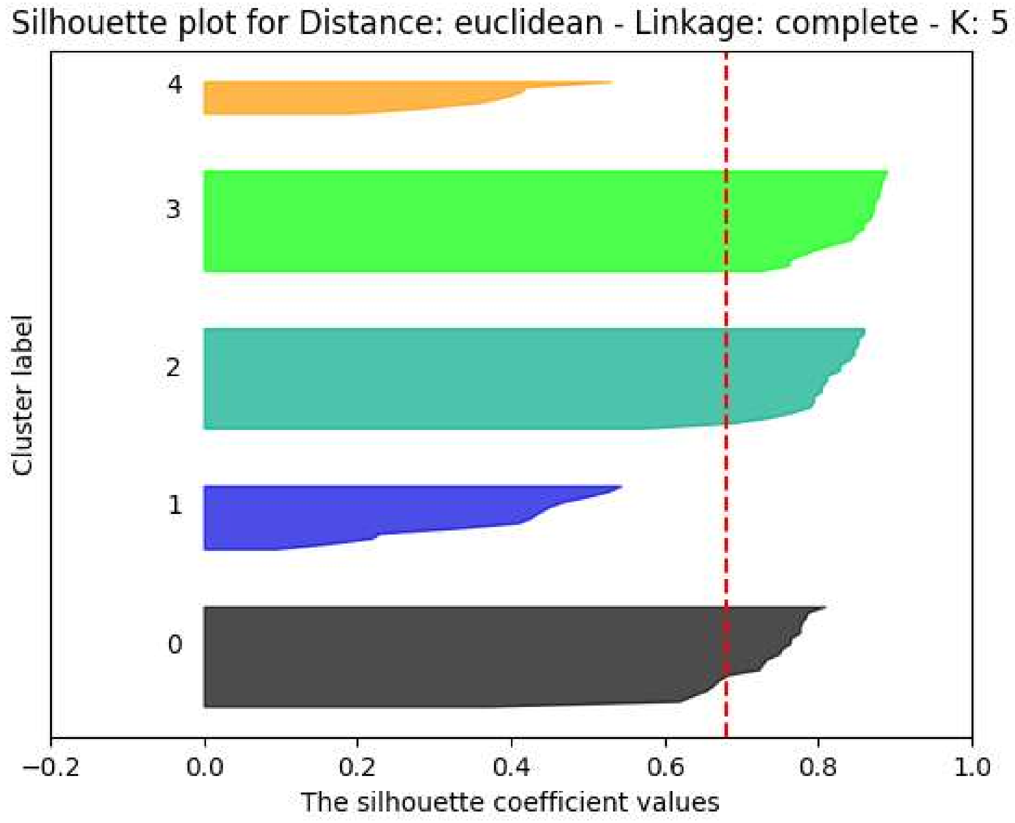
Comments: This shows good clustering.



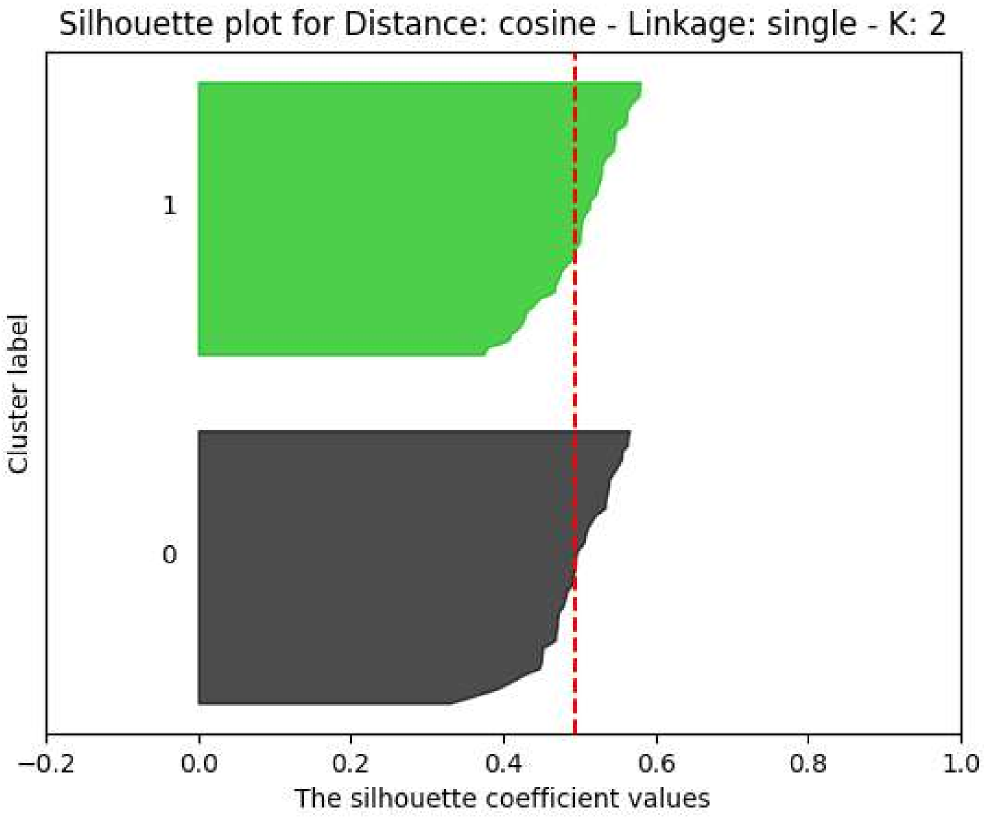
Comments: Label 1 has negative values, so some points are not clustered correctly.



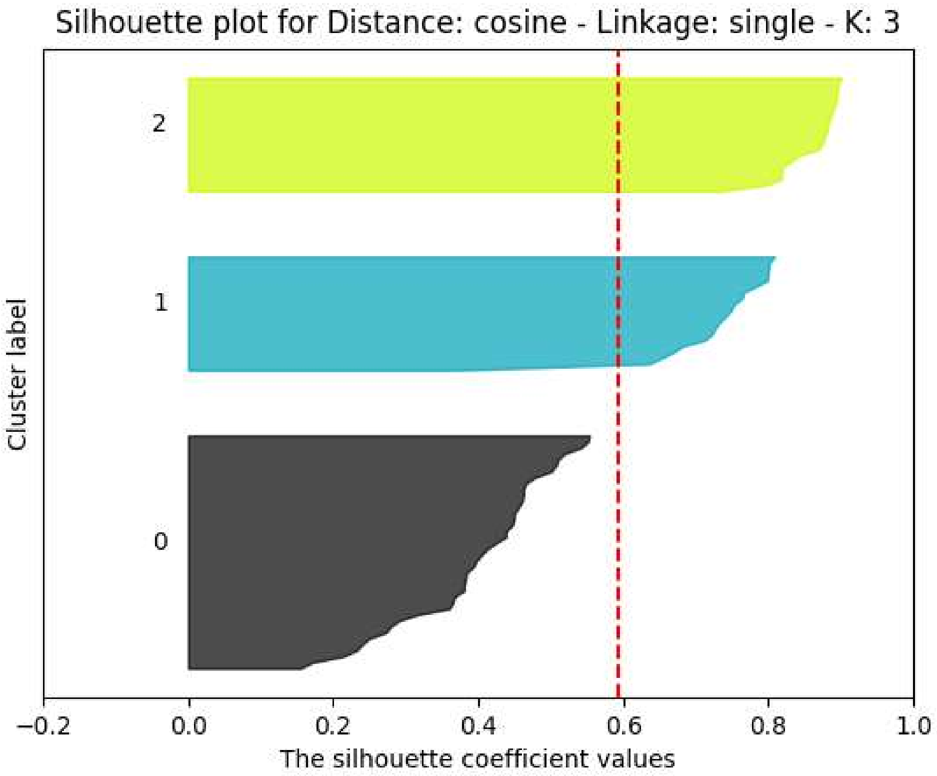
Comments: This shows good clustering.



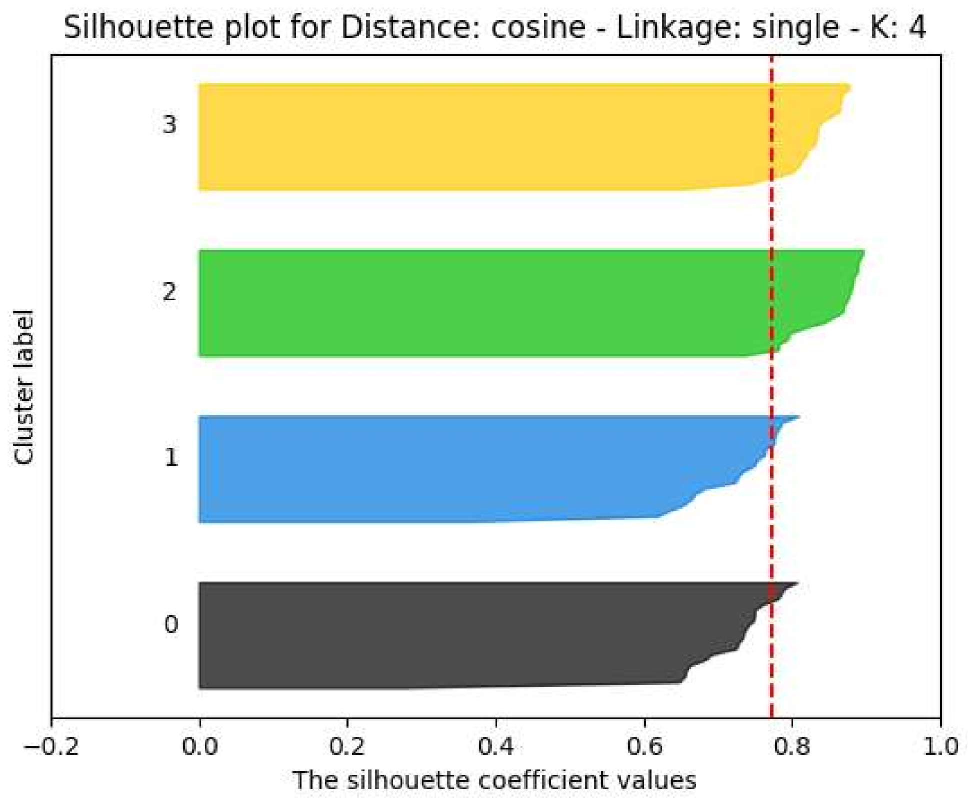
Comments: This shows good clustering, apart from Label 1 and 4 having very less data points.



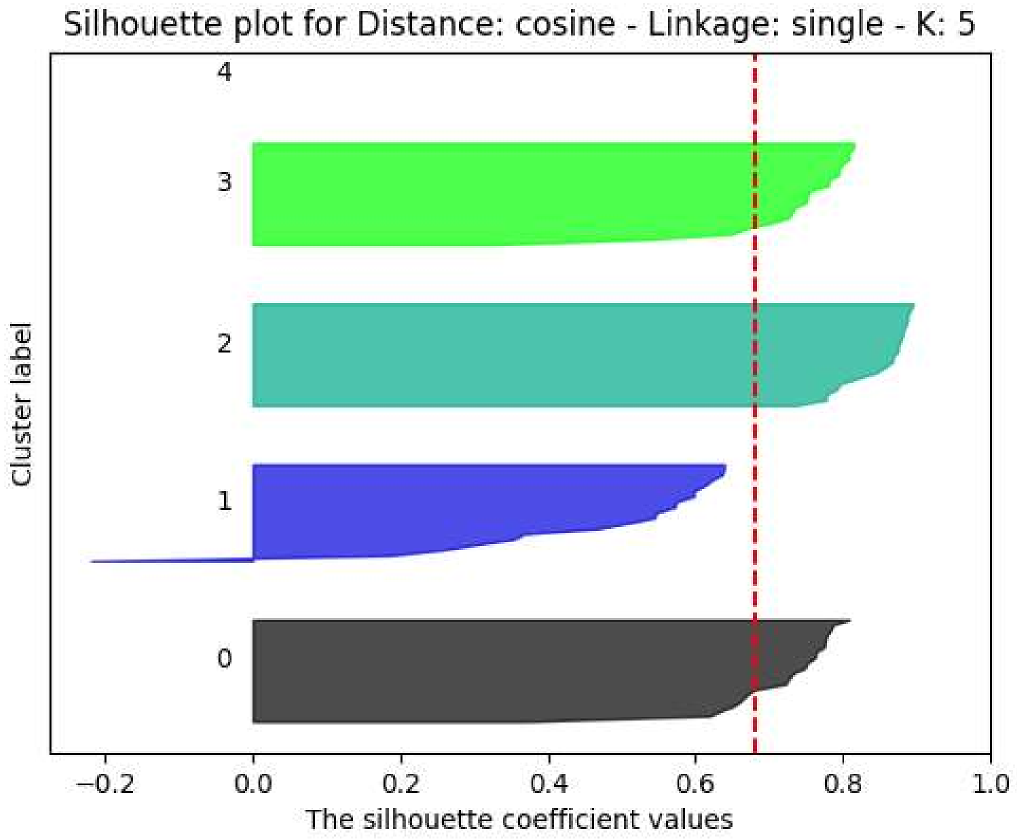
Comments: This shows good clustering.



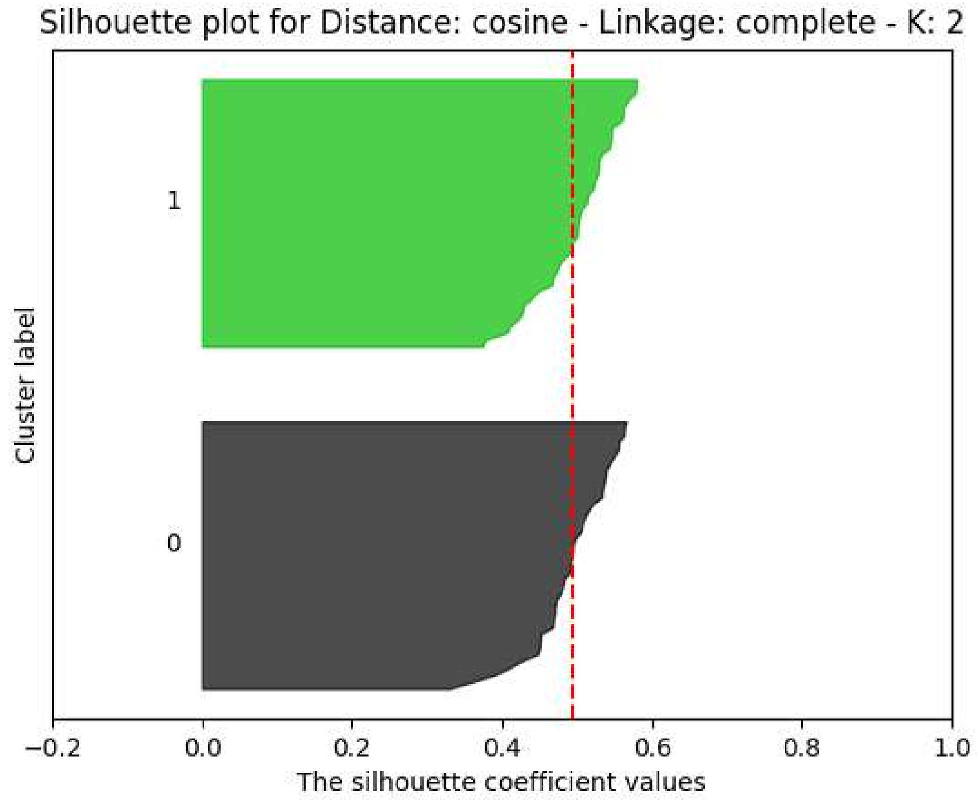
Comments: This shows good clustering.



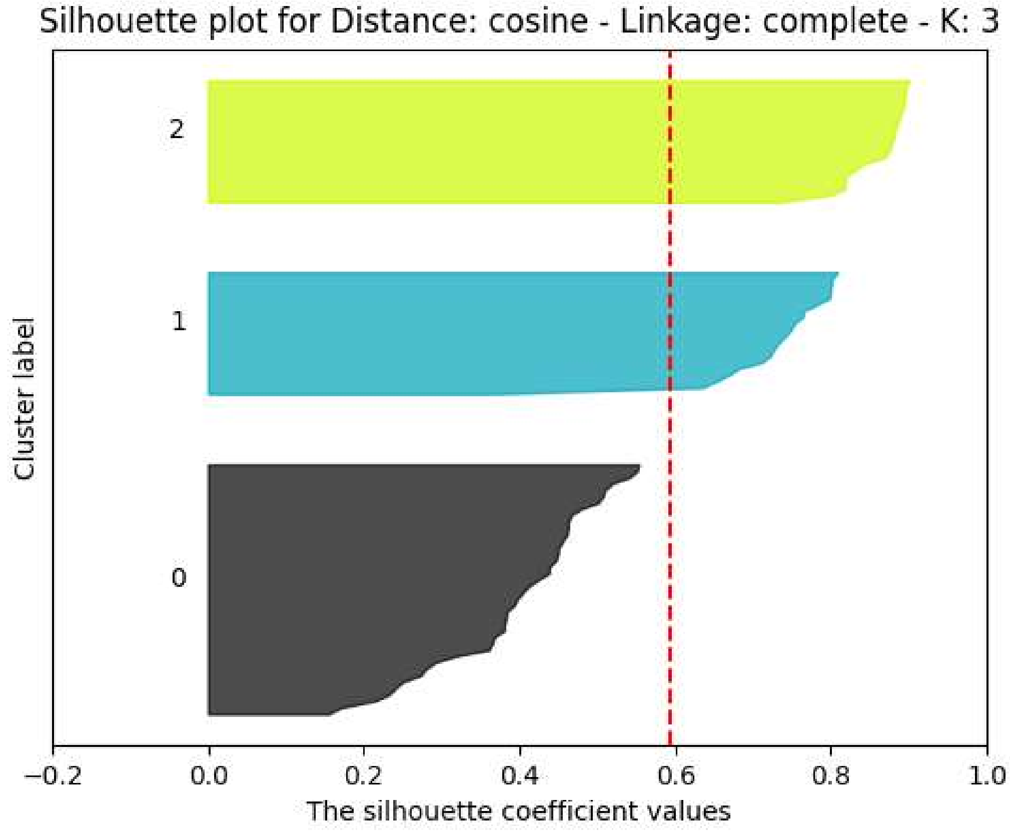
Comments: This shows good clustering.



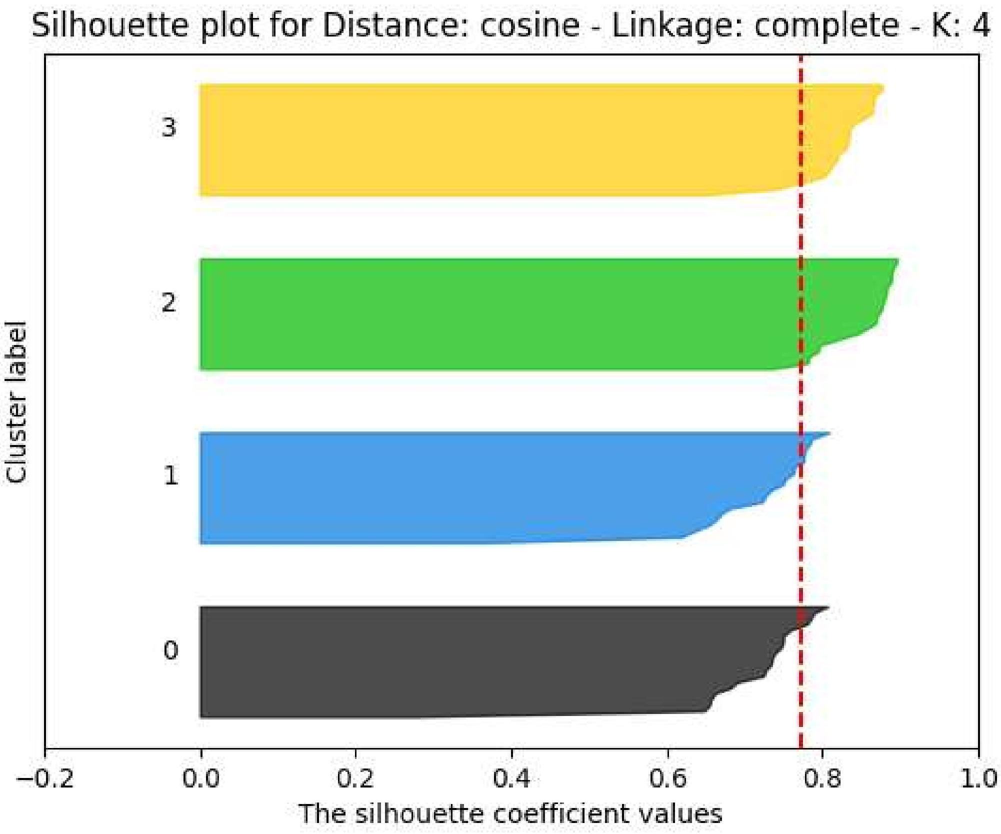
Comments: Label 1 has negative values, so some points are not clustered correctly.



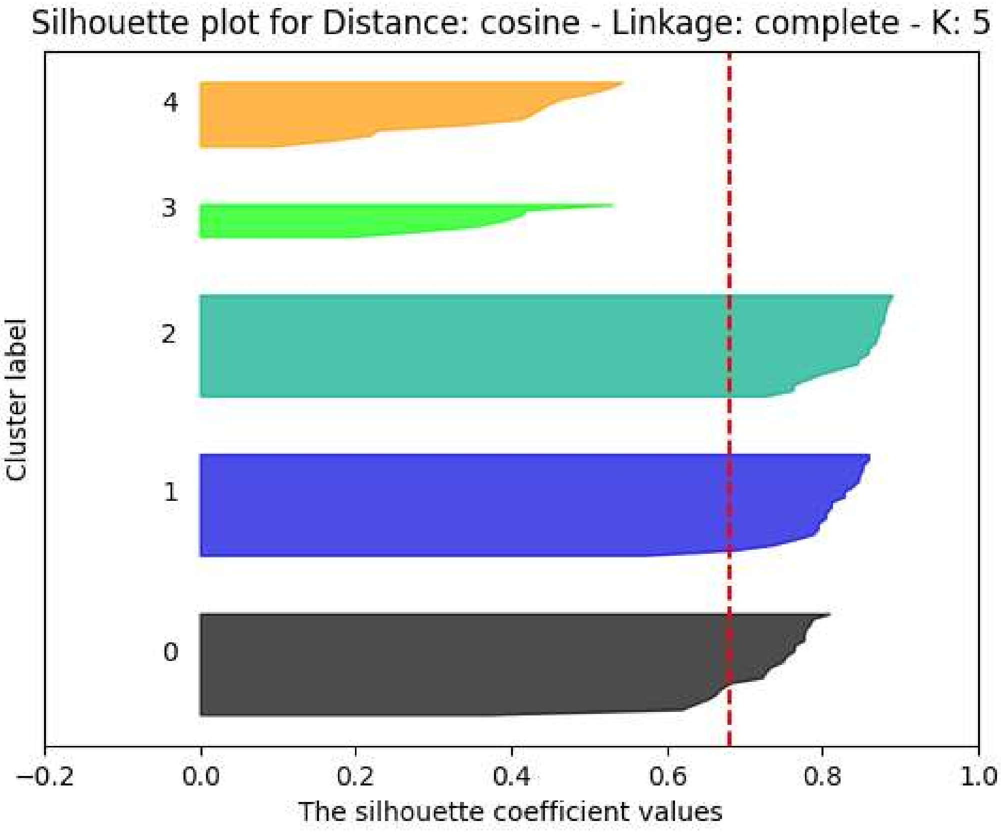
Comments: This shows good clustering.



Comments: This shows good clustering.



Comments: This shows good clustering.



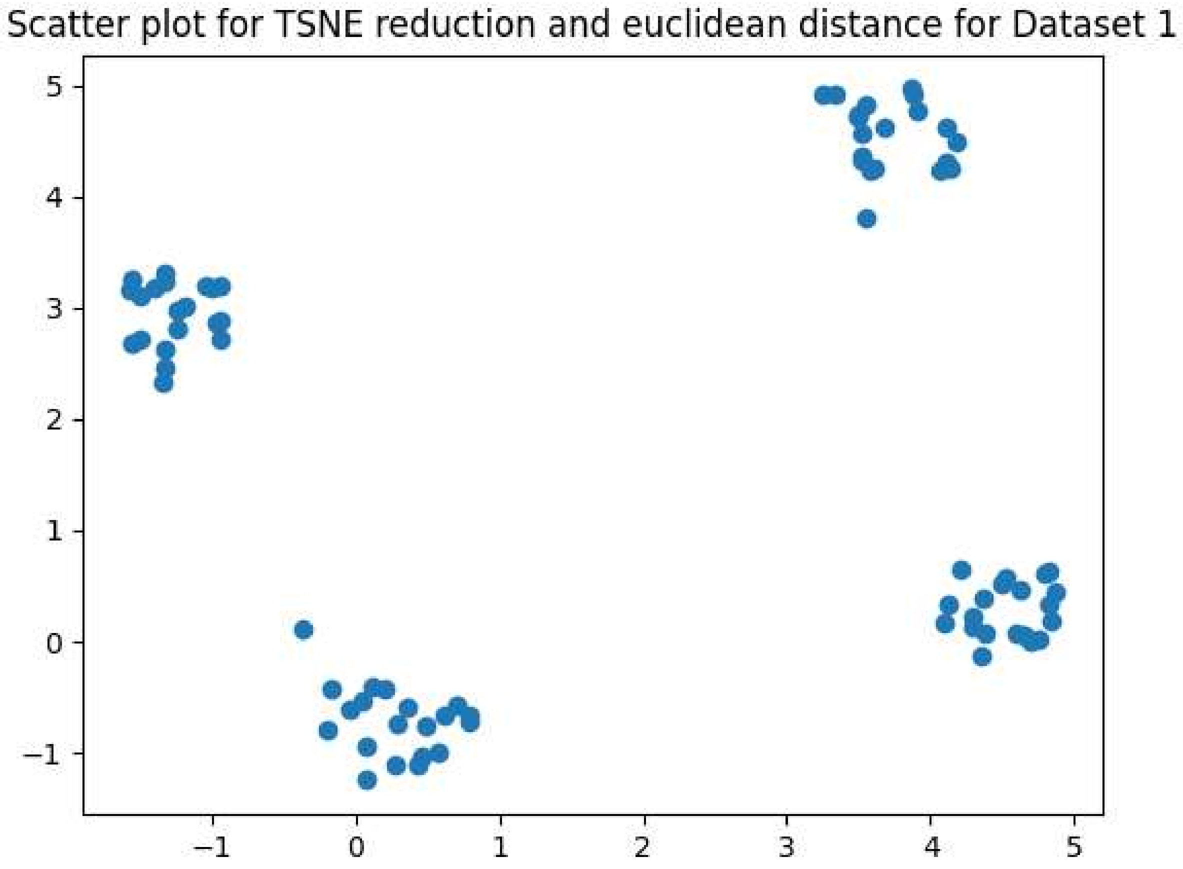
Comments: Label 3 and 4 has very few data points.

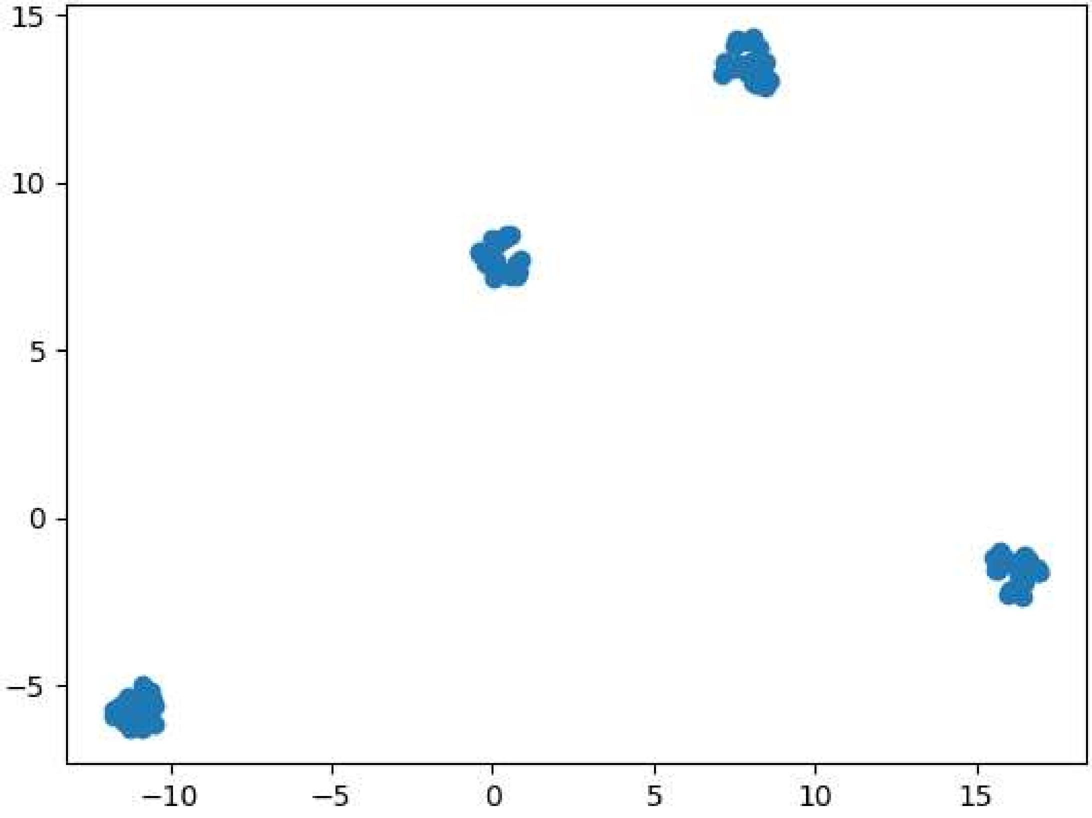
Best Result:

Distance: euclidean - Linkage: single - K: 4

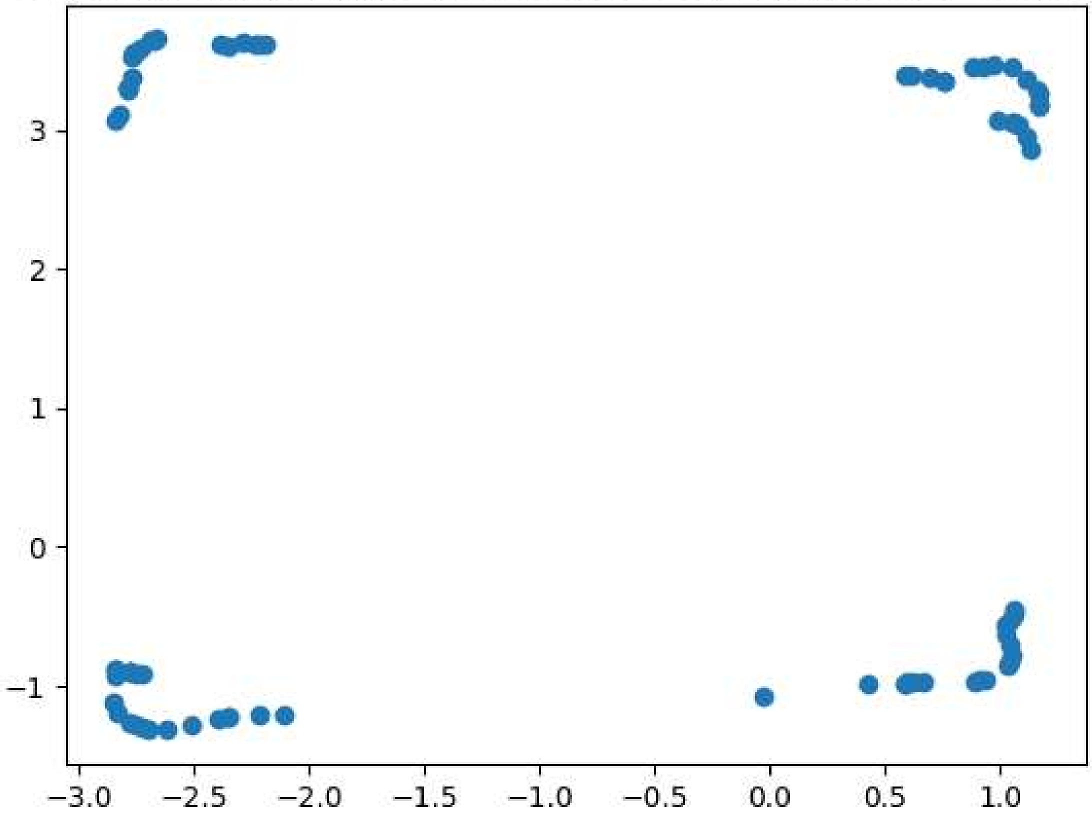
Average Silhouette Score: 0.77

Dimensionality Reduction:

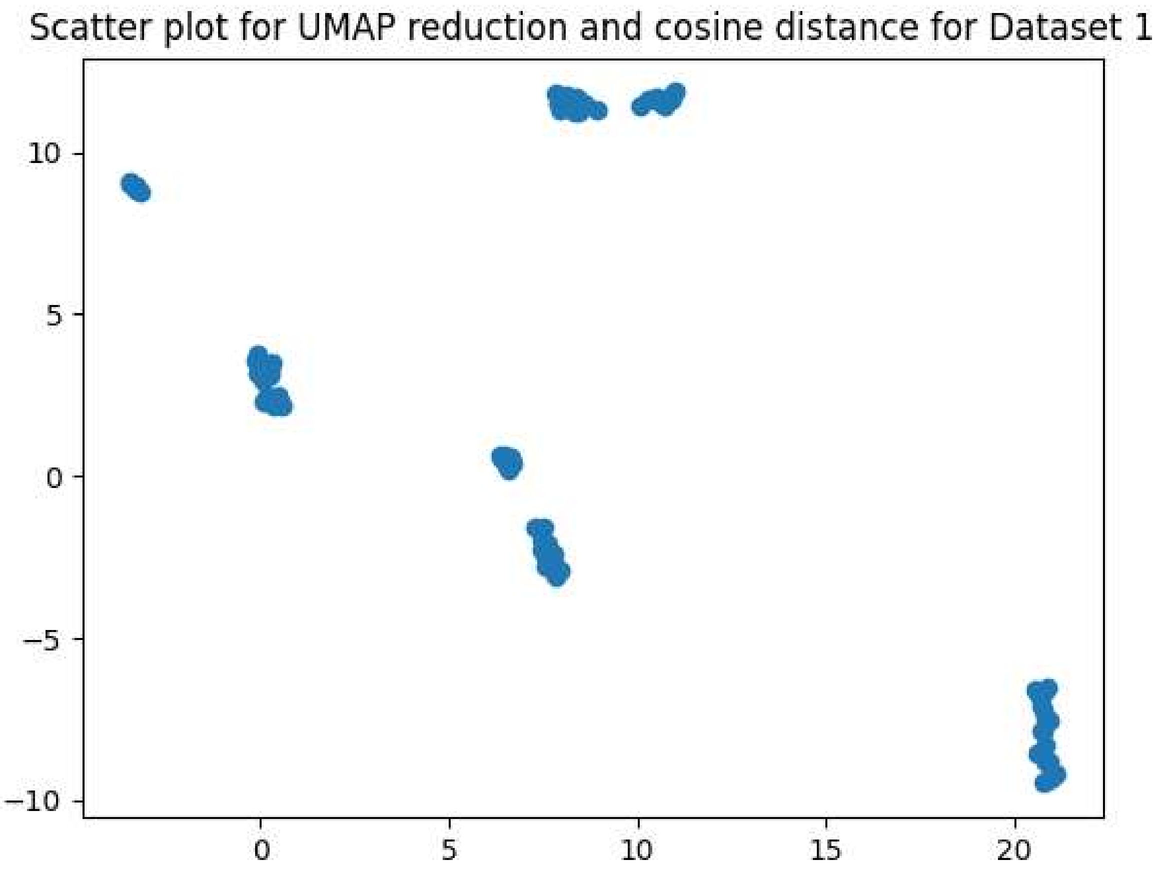




Scatter plot for UMAP reduction and euclidean distance for Dataset 1



Scatter plot for TSNE reduction and cosine distance for Dataset 1



Comments: All plots except UMAP with cosine show that the clusters should be 4. UMAP with cosine distance has some outliers, but for the rest In my opinion, the best one is TSNE with Euclidean distance as the points are not too much compressed, which shows that they have enough data and the clusters are compact.

Worst Case Analysis:

Worst-case run time analysis for HAC with respect to the number of data points (N) and dimension (D) of data instances.

For N points we need to find N-1 distances which becomes O(N^2D). After this we need to find the minimum distances for those, so it becomes **O(N^3 + N^2D).**