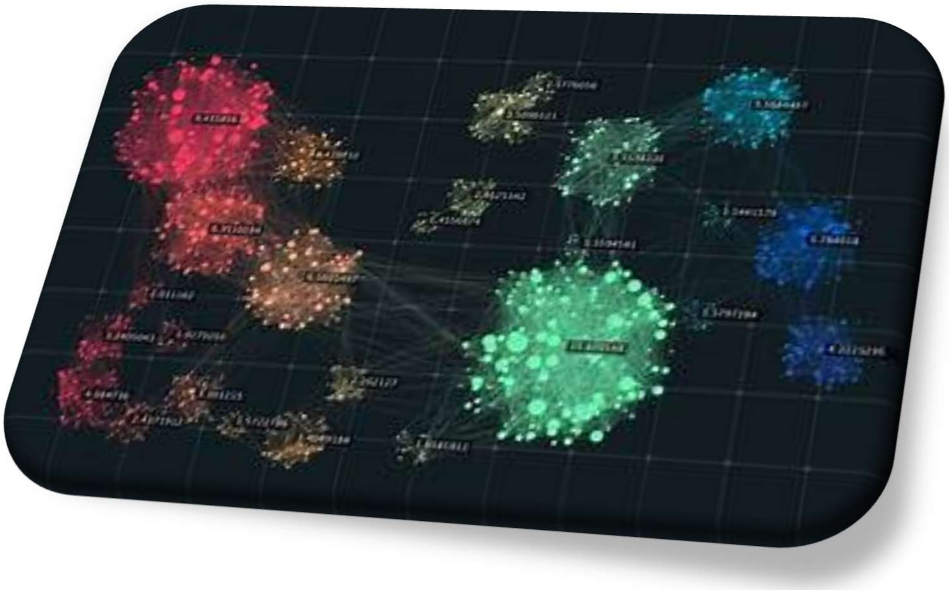


Data Clustering Analysis Report

Assignment # 2 Clustering

Mining of Massive Datasets BSCS 8A

Submitted to: Ms. Zareen Alamgir

Submitted by: Aisha Muhammad Nawaz (L200921)

Date Submitted: 31st March 2024 (Sunday)

National University of Computer and Emerging Sciences

Department of Computer Science

Lahore, Pakistan

INTRODUCTION

In this report, I will present the results of a data clustering analysis performed on datasets DS1 and DS2 provided on Google Classroom, using Apache PySpark. The objective of the analysis is to apply KMeans and Bisecting KMeans clustering algorithms to compare the performance of these algorithms (both built-in and self-implemented). The report provides insights into the clustering quality, optimal K values, clustering errors, and comparisons between the different algorithms and their results on different datasets.

METHODOLOGY

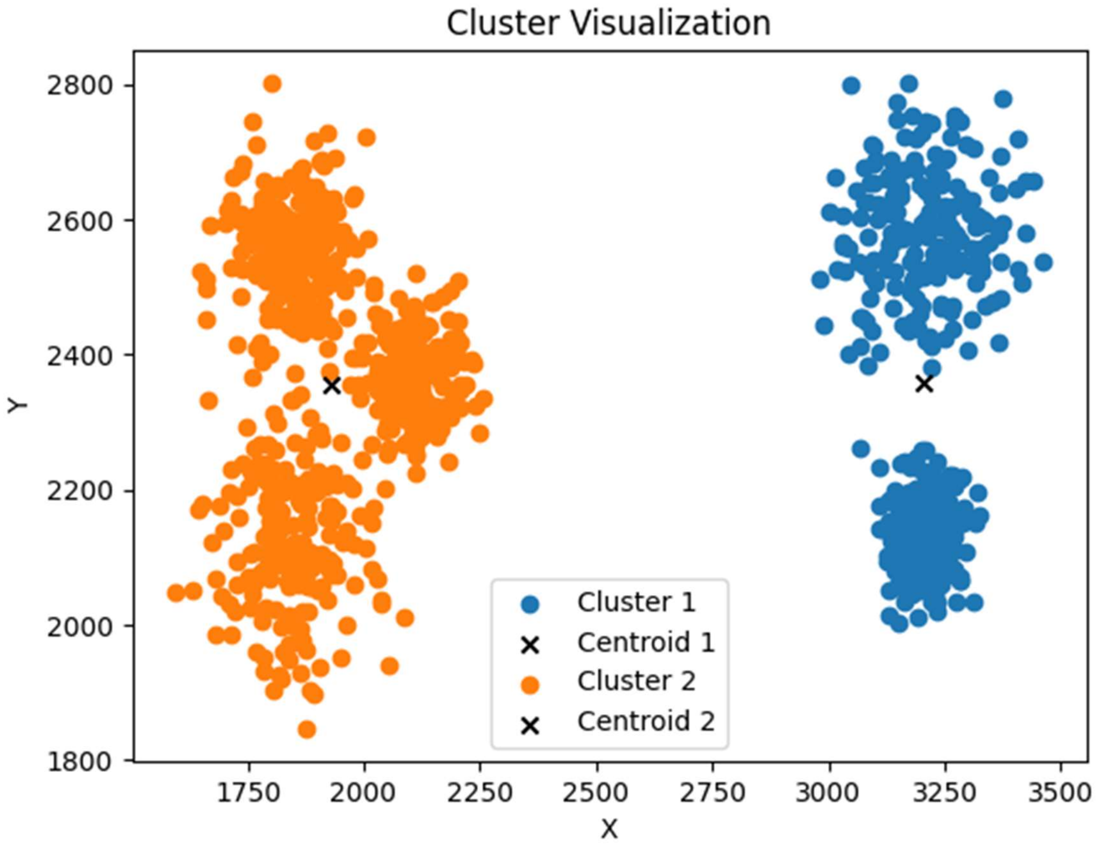
I wrote codes in PySpark as well as utilized Apache PySpark's machine learning library to implement KMeans and Bisecting KMeans clustering algorithms. The methodology involved the following steps:

1. Data Preprocessing: Parsing and assembling features from the raw data.
2. Running KMeans for different values of K.
3. Evaluating KMeans clustering quality using WSSE, BSSE, and Silhouette Coefficient.
4. Performing Bisecting KMeans Clustering for different values of K.
5. Assessing Bisecting KMeans Clustering quality using WSSE, BSSE, and Silhouette Coefficient.
6. Employing post-processing techniques to refine clustering results.

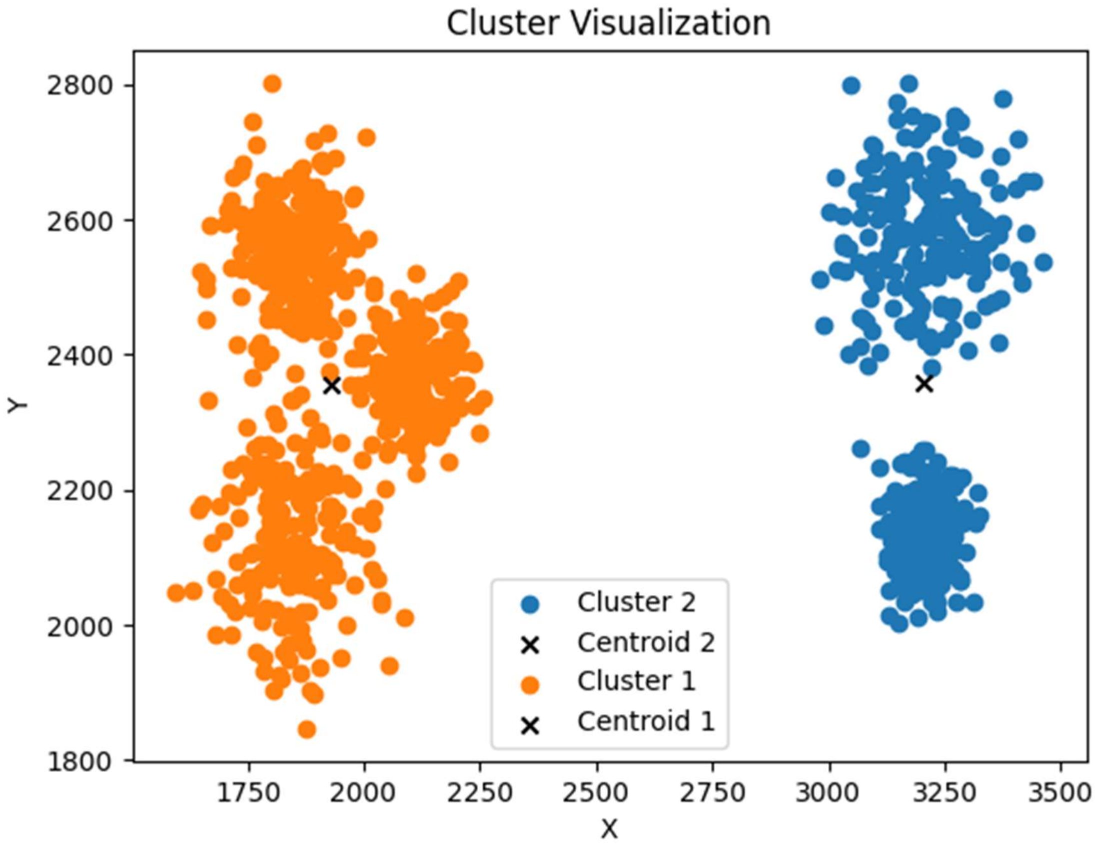
RESULTS

KMeans Clustering on DS1 (Without Using Built-in):

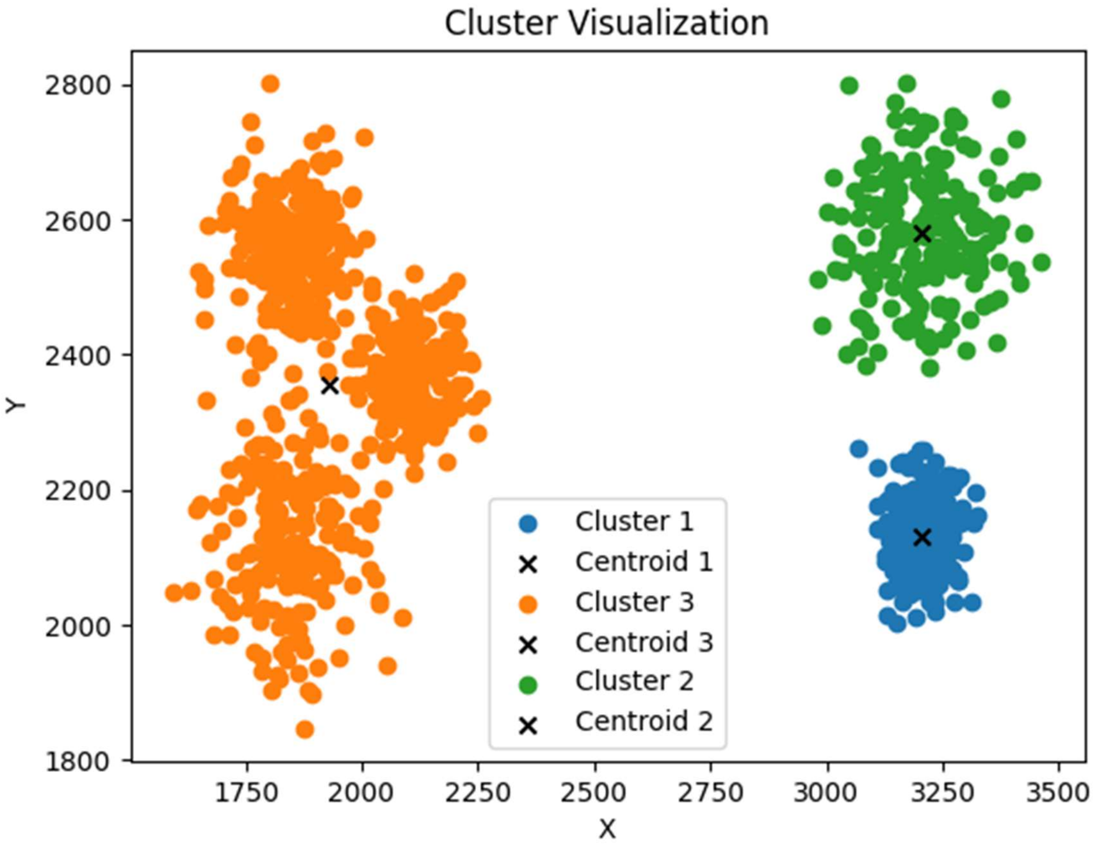
```
Running K Means for k= 2 Run #  1
Updated Centres:
[array([3204.11311054,2359.28791774]),
array([1928.50081833,2355.99181669])]

Error in each cluster:
Cluster  1 : WSSE:  24123380.77634961
Cluster  2 : WSSE:  35846847.7086743
BSSE = 406799396.0793503
SC: 0.7678397057752484

Time Taken to Converge: 4.086674451828003 seconds
Iterations Required to Converge: 3
```



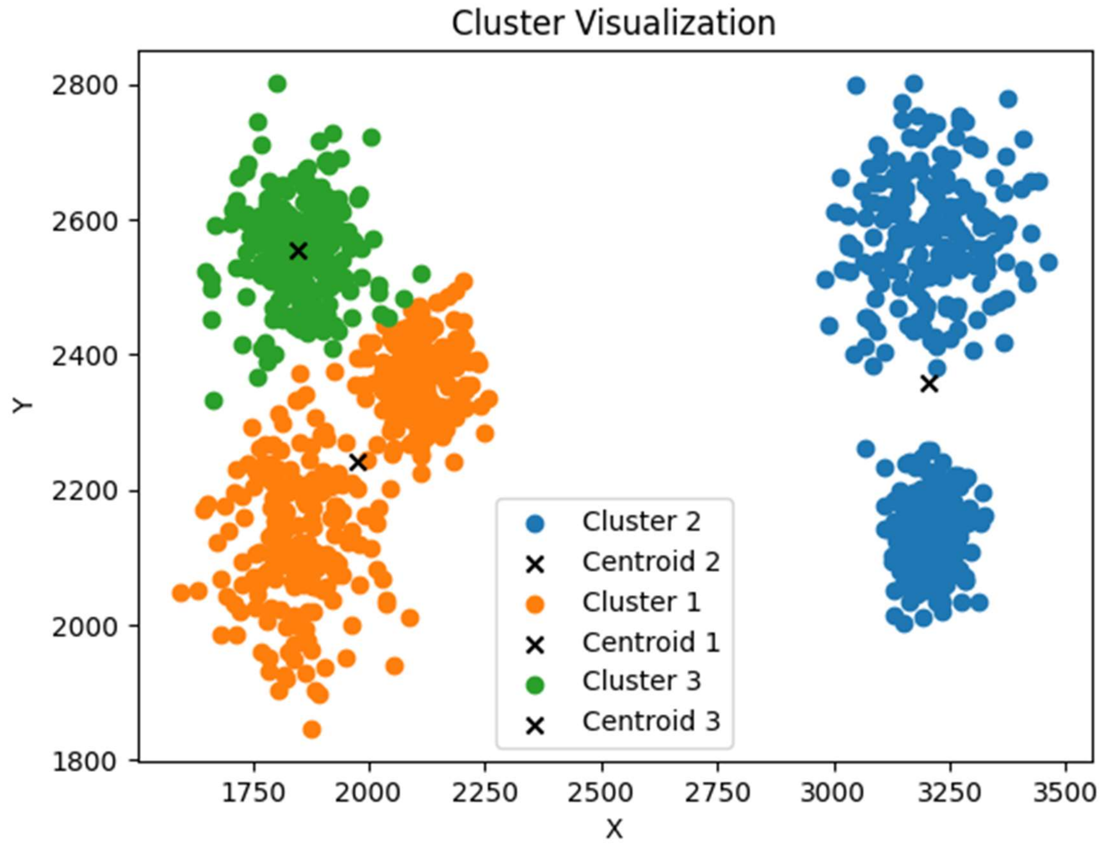
```
Running K Means for k= 2 Run #  2
Updated Centres :
 [array([1928.50081833, 2355.99181669]), array([3204.11311054,
2359.28791774])]

Error in each cluster:
WSSE =
Cluster  1 : WSSE:  35846847.7086743
Cluster  2 : WSSE:  24123380.77634961
BSSE = 406799396.0793503
SC: 0.7678397057752484

Time Taken to Converge: 1.7526657581329346 seconds

Iterations Required to Converge: 3
```



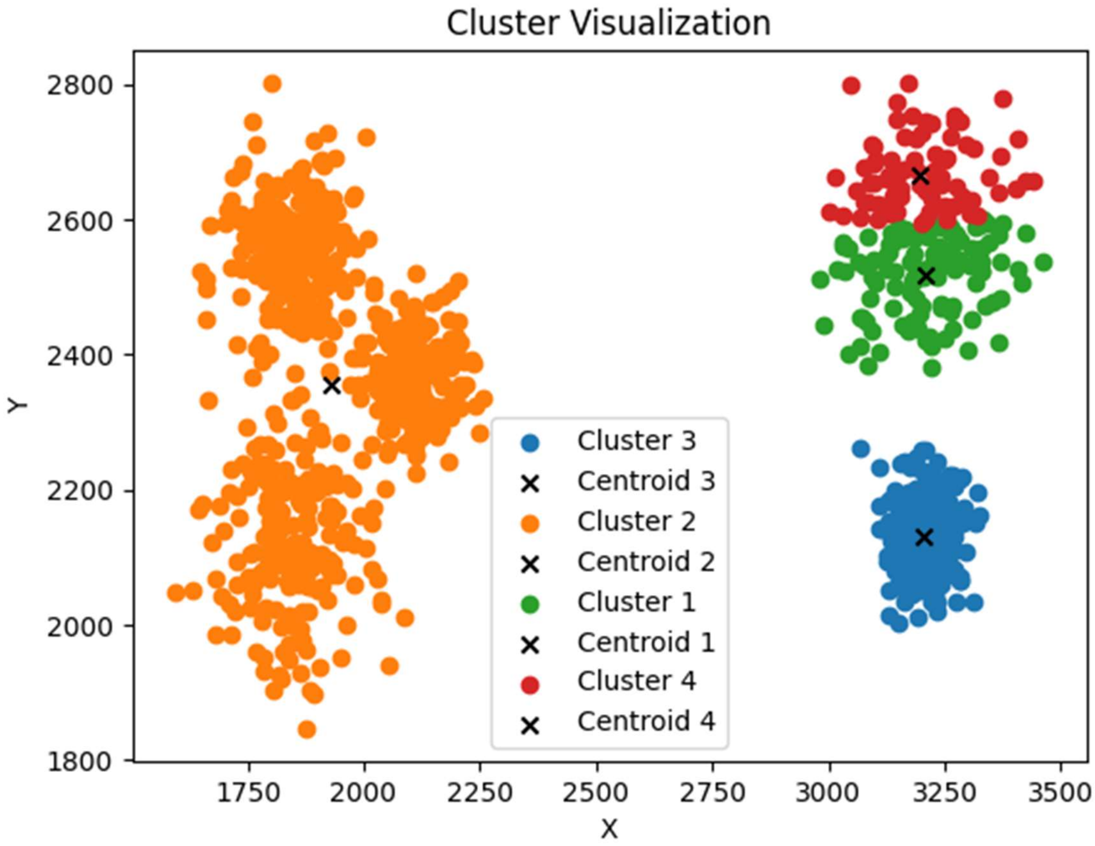
```
Running K Means for k= 3 Run #  1
Updated Centres :
 [array([3203.57068063, 2131.4921466 ]), array([3204.63636364,
2579.03030303]), array([1928.50081833, 2355.99181669])]

Error in each cluster:
WSSE =
Cluster  1 : WSSE:  1014080.5340314135
Cluster  3 : WSSE:  35846847.7086743
Cluster  2 : WSSE:  3637255.6363636362
BSSE = 531675448.62554723
SC: 0.8358596538337197

Time Taken to Converge: 2.2867202758789062 seconds

Iterations Required to Converge: 4
```



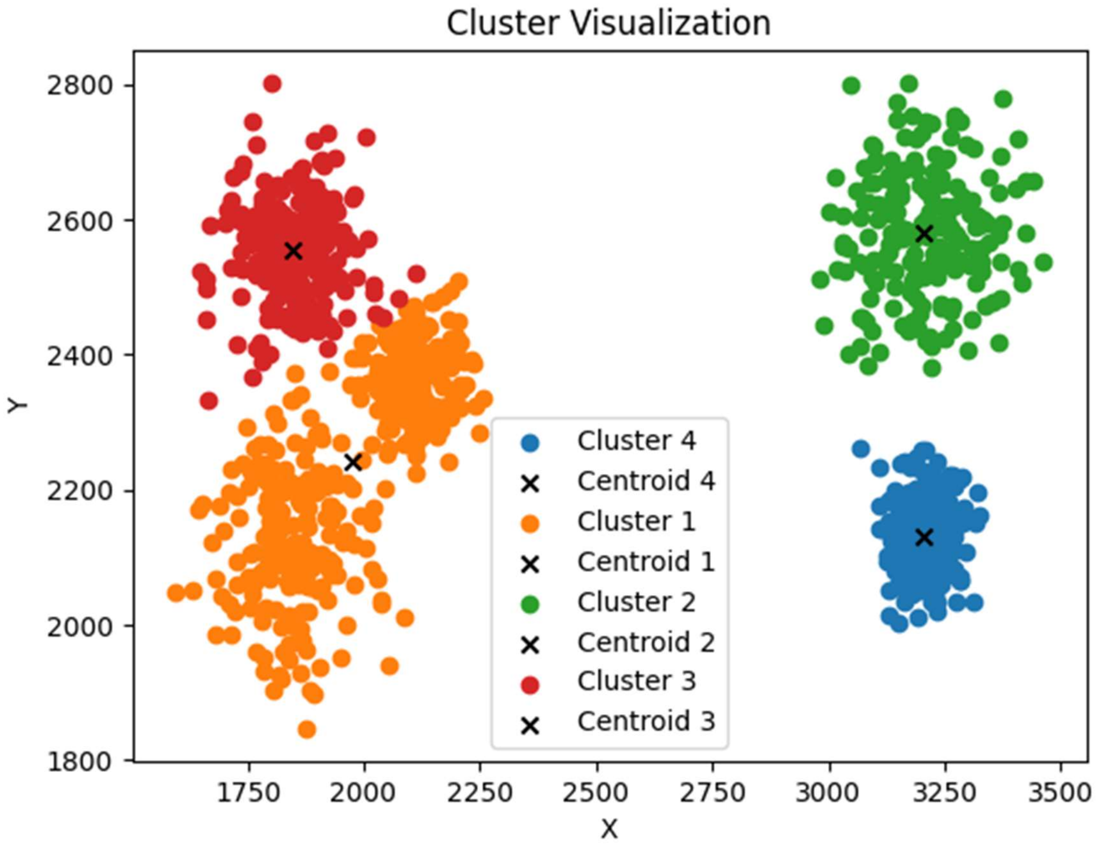
```
Running K Means for k= 3 Run #  2
Updated Centres :
 [array([1975.68134715, 2240.44041451]), array([3204.11311054,
2359.28791774]), array([1847.56      , 2554.22666667])]

Error in each cluster:
WSSE =
Cluster  1 : WSSE:  16900866.93523316
Cluster  3 : WSSE:  2616916.88
Cluster  2 : WSSE:  24123380.77634961
BSSE = 410596416.37526965
SC: 0.7864128816231716

Time Taken to Converge: 2.4949705600738525 seconds

Iterations Required to Converge: 4
```



```
Running K Means for k= 4 Run #  1
Updated Centres :
 [array([3209.14655172, 2517.26724138]), array([1928.50081833,
2355.99181669]), array([3203.57068063, 2131.4921466 ]),
array([3198.25609756, 2666.40243902])]

Error in each cluster:
WSSE =
Cluster  3 : WSSE:  1014080.5340314135
Cluster  1 : WSSE:  1557467.2241379314
Cluster  2 : WSSE:  35846847.7086743
Cluster  4 : WSSE:  1005609.3414634147
BSSE = 622701110.2269468
SC: 0.8309420421493645

Time Taken to Converge: 3.7237966060638428 seconds

Iterations Required to Converge: 6
```



```
Running K Means for k= 4 Run #  2
Updated Centres :
 [array([1975.68134715, 2240.44041451]), array([3204.63636364,
2579.03030303]), array([1847.56      , 2554.22666667]),
array([3203.57068063, 2131.4921466 ])]

Error in each cluster:
WSSE =
Cluster  1 : WSSE:  16900866.93523316
Cluster  3 : WSSE:  2616916.88
Cluster  4 : WSSE:  1014080.5340314135
Cluster  2 : WSSE:  3637255.6363636362
BSSE = 440642108.193802
SC: 0.8574974235626869

Time Taken to Converge: 5.285777807235718 seconds

Iterations Required to Converge: 6
```



Overall Result

Metric	K=2, Run=1	K=2, Run=2	K=3, Run=1	K=3, Run=2	K=4, Run=1	K=4, Run=2
WSSE	59970228.49	59970228.49	40498183.88	43641164.59	**1005609.34**	24169119.99
BSSE	**406799396.08**	**406799396.08**	531675448.63	410596416.38	622701110.23	440642108.19
SC	0.76784	0.76784	0.83586	0.78641	0.83094	**0.85750**
Time Taken	4.087 secs	**1.754 secs**	2.287 secs	2.495 secs	3.724 secs	5.286 secs
Iterations	**3**	**3**	4	4	6	6

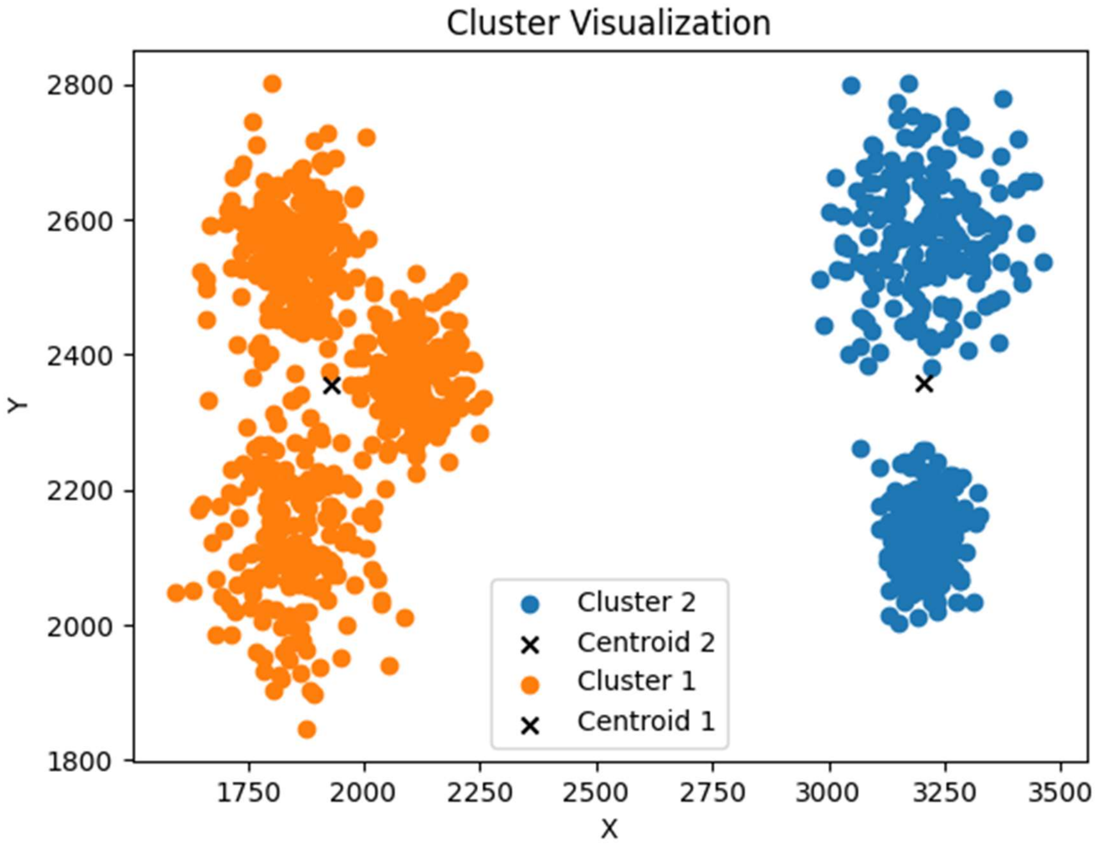
Based on the silhouette coefficient and relatively low mistakes**, K=2** produces the most balanced clustering solution, according to the data. It converges in 1.754 secs while only taking 3 iterations to do so. The SC value is 0.76784, WSSE is 59970228.49 and BSSE is 406799396.08. Higher K values, however, are generally observed to result in longer convergence periods and a greater number of iterations needed to reach convergence, indicating an increase in computational complexity. Overall, the results indicate that for the given dataset, K=2 offers a good compromise between computing efficiency and clustering accuracy but the highest SC value is observed when K=4 (SC was around 0.85750)

Bisecting KMeans Clustering on DS1 (without using built-in):

```
Running K Means for k= 2 Run #  1
Updated Centres :
 [array([1928.50081833, 2355.99181669]), array([3204.11311054,
2359.28791774])]
```

```
Error in each cluster:
WSSE =
Cluster  1 : WSSE:  35846847.7086743
Cluster  2 : WSSE:  24123380.77634961
BSSE = 406799396.0793503
SC: 0.7678397057752484

Time Taken to Converge: 1.8573393821716309 seconds

Iterations Required to Converge: 3
Running K Means for k= 2 Run #  2
Updated Centres :
 [array([3204.11311054, 2359.28791774]), array([1928.50081833,
2355.99181669])]

Error in each cluster:
WSSE =
Cluster  1 : WSSE:  24123380.77634961
Cluster  2 : WSSE:  35846847.7086743
BSSE = 406799396.0793503
SC: 0.7678397057752484

Time Taken to Converge: 1.9847311973571777 seconds

Iterations Required to Converge: 2
Running K Means for k= 2 Run #  3
Updated Centres :
 [array([1928.50081833, 2355.99181669]), array([3204.11311054,
2359.28791774])]

Error in each cluster:
WSSE =
Cluster  1 : WSSE:  35846847.7086743
Cluster  2 : WSSE:  24123380.77634961
BSSE = 406799396.0793503
SC: 0.7678397057752484

Time Taken to Converge: 1.8748741149902344 seconds

Iterations Required to Converge: 3
Running K Means for k= 2 Run #  4
Updated Centres :
 [array([3204.11311054, 2359.28791774]), array([1928.50081833,
2355.99181669])]

Error in each cluster:
WSSE =
Cluster  1 : WSSE:  24123380.77634961
Cluster  2 : WSSE:  35846847.7086743
BSSE = 406799396.0793503
SC: 0.7678397057752484

Time Taken to Converge: 1.2237849235534668 seconds

Iterations Required to Converge: 2
Running K Means for k= 2 Run #  5
```

```
Updated Centres :
 [array([3204.11311054, 2359.28791774]), array([1928.50081833,
2355.99181669])]

Error in each cluster:
WSSE =
Cluster  1 : WSSE:  24123380.77634961
Cluster  2 : WSSE:  35846847.7086743
BSSE = 406799396.0793503
SC: 0.7678397057752484

Time Taken to Converge: 1.750960350036621 seconds

Iterations Required to Converge: 2
Running K Means for k= 2 Run #  6
Updated Centres :
 [array([1928.50081833, 2355.99181669]), array([3204.11311054,
2359.28791774])]

Error in each cluster:
WSSE =
Cluster  1 : WSSE:  35846847.7086743
Cluster  2 : WSSE:  24123380.77634961
BSSE = 406799396.0793503
SC: 0.7678397057752484

Time Taken to Converge: 1.2119860649108887 seconds

Iterations Required to Converge: 2
Cluster 1 Centroid: [array([3204.11311054, 2359.28791774]),
array([1928.50081833, 2355.99181669])]
Cluster 2 Centroid: [array([3204.11311054, 2359.28791774]),
array([1928.50081833, 2355.99181669])]
Cluster 3 Centroid: [array([1928.50081833, 2355.99181669]),
array([3204.11311054, 2359.28791774])]
```

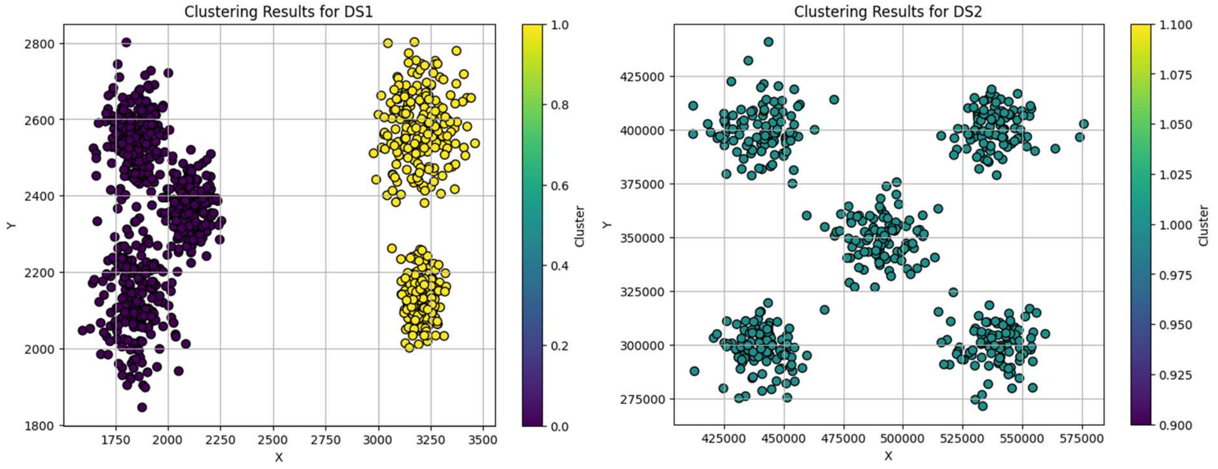


The dataset DS1 with K=2 was subjected to many runs of the Bisecting KMeans clustering method in order to evaluate clustering consistency and convergence behavior. The method took two to three iterations to converge across various runs, and each run took between one and two seconds to finish. The centroids of both clusters converged to comparable locations, indicating that the clustering results were consistent between runs. Stable clustering solutions were indicated by the consistent within-cluster sum of squared errors (WSSE) and between-cluster sum of square errors (BSSE) across runs. Additionally, the silhouette coefficient (SC) held steady at approximately 0.77, suggesting a good distance between clusters. In splitting the dataset into two separate clusters, the Bisecting KMeans algorithm showed stability and consistency overall.

KMeans Clustering on DS1& DS2 (Using Built-in):

```
Silhouette score for K=2: 0.9300334351025269, WSSE: 59970228.48502394,
BSSE: 3254395.168634804
Silhouette score for K=3: 0.9076503023053574, WSSE: 40498183.87906939,
BSSE: 7109525.778275301
Silhouette score for K=4: 0.8381558342151313, WSSE: 39183996.01079804,
BSSE: 10991742.558920657
Optimal K: 2, Silhouette Score: 0.9300334351025269, Best WSSE:
59970228.48502394
Silhouette score for K=2: 0.977689319886371, WSSE: 7011360115400.669,
BSSE: 192494316366.1994
Silhouette score for K=3: 0.9429828430500716, WSSE: 6164096499747.533,
BSSE: 397700954255.82446
Silhouette score for K=4: 0.707571350834473, WSSE: 3404478608094.0596,
BSSE: 948884433265.5885
```

```
Optimal K: 2, Silhouette Score: 0.977689319886371, Best WSSE:
7011360115400.669
```



Overall Results for Dataset 1

Metric	K=2	K=3	K=4
WSSE	59970228.49	40498183.88	**39183996.01**
BSSE	**3254395.17**	7109525.78	10991742.56
SC	**0.930**	0.908	0.838

Overall Results for Dataset 2

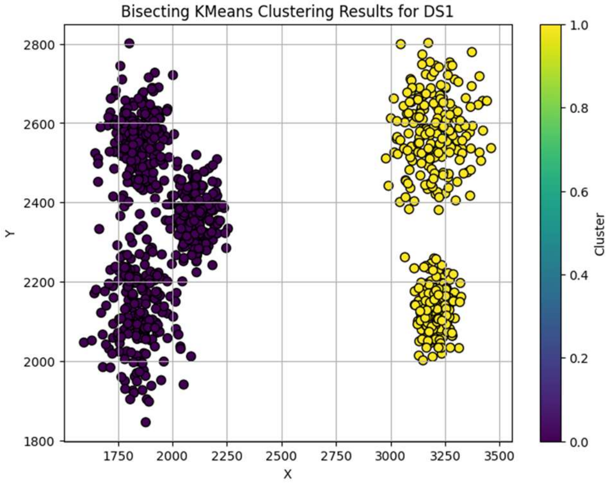
Metric	K=2	K=3	K=4
WSSE	7011360115400.67	6164096499747.53	**3404478608094.06**
BSSE	**192494316366.20**	397700954255.82	948884433265.59
SC	**0.978**	0.943	0.708

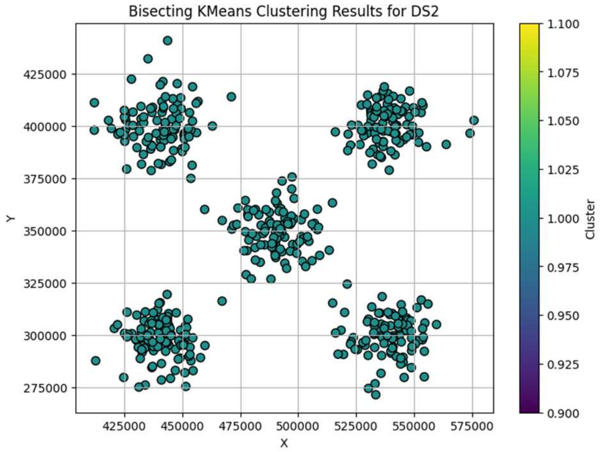
The K-means clustering algorithm was applied to datasets DS1 and DS2 using the built-in function, considering different values of K (2, 3, and 4). For DS1, K=2 emerged as the optimal choice, with a silhouette score of 0.930 and a WSSE of 59,970,228.49. Similarly, for DS2, K=2 was again identified as the optimal value, with a higher silhouette score of 0.978 and a larger WSSE of 7,011,360,115,400.67. Despite the larger WSSE in DS2, the higher silhouette score indicates better cluster separation. These findings suggest that K=2 provides the most effective clustering solution for both datasets using the K-means algorithm with the built-in function.

Bisecting KMeans Clustering on DS1 & DS2 (Using Built-in):

```
Silhouette score for K=2: 0.9300334351025269, WSSE: 59970228.48502394,
BSSE: 3254395.168634804
Silhouette score for K=3: 0.759066658536799, WSSE: 41954375.7563119, BSSE:
7165062.506195096
Silhouette score for K=4: 0.561342256722003, WSSE: 31027111.544380452,
BSSE: 13331501.010269053
Optimal K: 2, Silhouette Score: 0.9300334351025269, Best WSSE:
59970228.48502394
Silhouette score for K=2: 0.977689319886371, WSSE: 7011360115400.669,
BSSE: 192494316366.1994
```

```
Silhouette score for K=3: 0.7460287620725604, WSSE: 4022682726386.8555,
BSSE: 410339859919.4301
Silhouette score for K=4: 0.7189355963485206, WSSE: 3131880783349.231,
BSSE: 844894681962.0858
Optimal K: 2, Silhouette Score: 0.977689319886371, Best WSSE:
7011360115400.669
```





Overall Results for Dataset 1

Metric	K=2	K=3	K=4
WSSE	59970228.49	41954375.76	**31027111.54**
BSSE	**3254395.17**	7165062.51	13331501.01
SC	**0.930**	0.759	0.561

Overall Results for Dataset 2

Metric	K=2	K=3	K=4
WSSE	7011360115400.67	4022682726386.85	**3131880783349.23**
BSSE	**192494316366.20**	410339859919.43	844894681962.09
SC	**0.978**	0.746	0.719

The results of the built-in Bisecting K-means algorithm for datasets DS1 and DS2 were analyzed based on silhouette scores and within-cluster sum of squared errors (WSSE) for different values of K (2, 3, and 4). For DS1, the silhouette score was highest for K=2, indicating better cluster cohesion and separation, with a score of 0.930 and a WSSE of 59,970,228.49. Similarly, for DS2, K=2 yielded the highest silhouette score of 0.978 and a WSSE of 7,011,360,115,400.67. Despite the larger WSSE in DS2, the higher silhouette score suggests improved clustering quality with K=2. These findings suggest that K=2 provides the most effective clustering solution for both datasets using the Bisecting K-means algorithm with the built-in function.

Comparing the clustering results of the K-means and Bisecting K-means for all the datasets:

Dataset 1			
K-Means			
Metric	K=2	K=3	K=4
WSSE	59970228.49	40498183.88	**39183996.01**
BSSE	**3254395.17**	7109525.78	10991742.56
SC	**0.930**	0.908	0.838
Bisecting K-Means			
Metric	K=2	K=3	K=4
WSSE	59970228.49	41954375.76	**31027111.54**
BSSE	**3254395.17**	7165062.51	13331501.01
SC	**0.930**	0.759	0.561

Dataset 2			
Metric	K=2	K=3	K=4
WSSE	7011360115400.67	6164096499747.53	**3404478608094.06**
BSSE	**192494316366.20**	397700954255.82	948884433265.59
SC	**0.978**	0.943	0.708

Bisecting K-means

Metric	K=2	K=3	K=4
WSSE	7011360115400.67	4022682726386.85	**3131880783349.23**
BSSE	**192494316366.20**	410339859919.43	844894681962.09
SC	**0.978**	0.746	0.719

Comparing the clustering results of K-means and Bisecting K-means across both datasets reveals nuanced performance variations. In Dataset 1, K-means and Bisecting K-means demonstrate similar within-cluster sum of squared errors (WSSE) for K=2, indicating consistent clustering quality. However, for higher K values, K-means generally exhibits lower WSSE and between-cluster sum of squared errors (BSSE), suggesting improved cluster separation. Conversely, Bisecting K-means shows a sharper decline in silhouette coefficient (SC) with increasing K, indicating decreased cluster cohesion and separation beyond K=2.

In Dataset 2, K-means and Bisecting K-means yield comparable WSSE, SC & BSSE for K=2. Both algorithms show a decrease in SC with higher K values, suggesting diminished clustering effectiveness. Despite this, Bisecting K-means displays a more pronounced decline in SC, especially noticeable for K=3 and K=4. This is similar to what was observed in Dataset 1,

Overall, while K-means generally outperforms Bisecting K-means in clustering quality across both datasets, particularly evident with higher K values, Bisecting K-means offers competitive results for K=2 clusters in Dataset 1.

CONCLUSION

The comparison between built-in and self-implemented K-means clustering algorithms, alongside the integration of Bisecting K-means, underscores several key points. Built-in algorithms excel in efficiency, boasting faster convergence times and lower computational complexity, likely due to optimization techniques. However, self-implemented methods offer flexibility and insight into algorithmic intricacies. Both approaches yield comparable clustering accuracy, with K=2 consistently emerging as the optimal choice for the datasets examined. The stability and reliability of clustering results are evident across different runs, particularly with Bisecting K-means. Notably, while K=2 remains optimal, the exploration of higher K values unveils potential enhancements in cluster separation, albeit at increased computational cost. Ultimately, the selection between built-in and self-implemented methods hinges on the balance between computational efficiency and customization needs, with Bisecting K-means offering robustness for datasets requiring consistent clustering solutions.