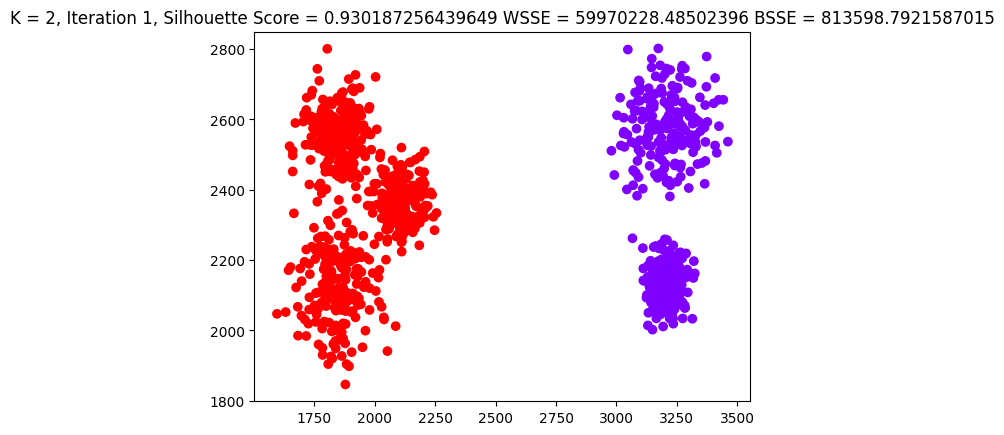
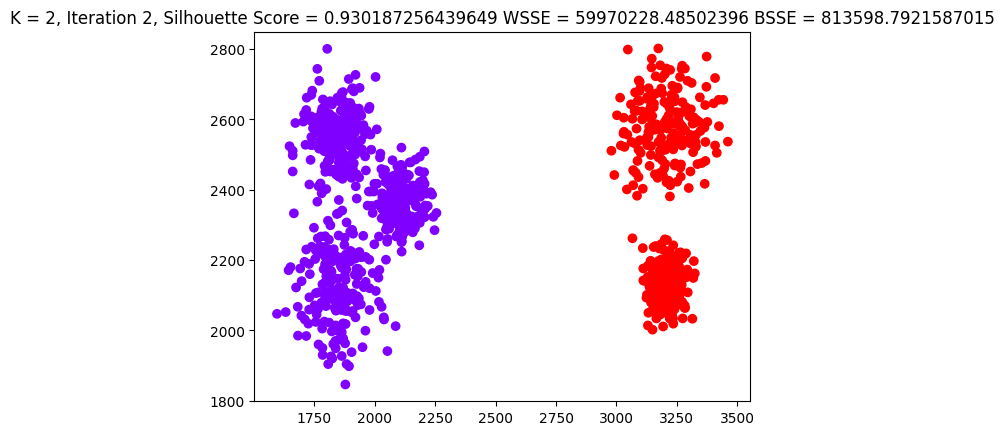
**Clustering Report**

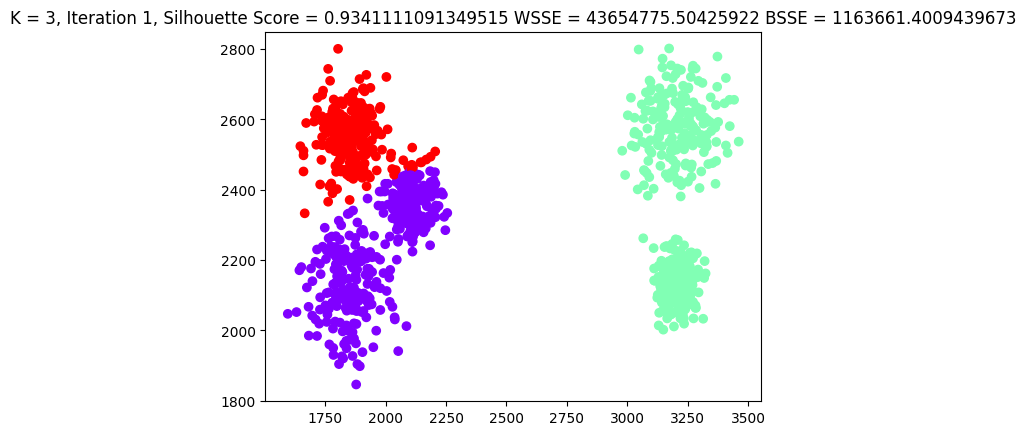
DS1.txt – Kmeans PYSPARK code given in class

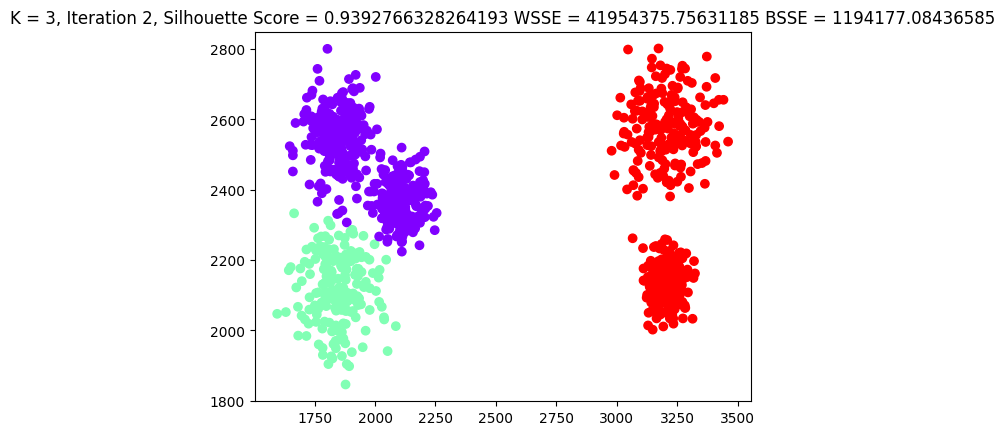
Running the code given in class for Kmeans clustering on DS1.txt yielded *K = 9* as the best number of clusters when clusters were tested by taking 2 to 9 centroids (quality metric = Silhouette Score*. K = 9* had the best score which was **0.983689**).

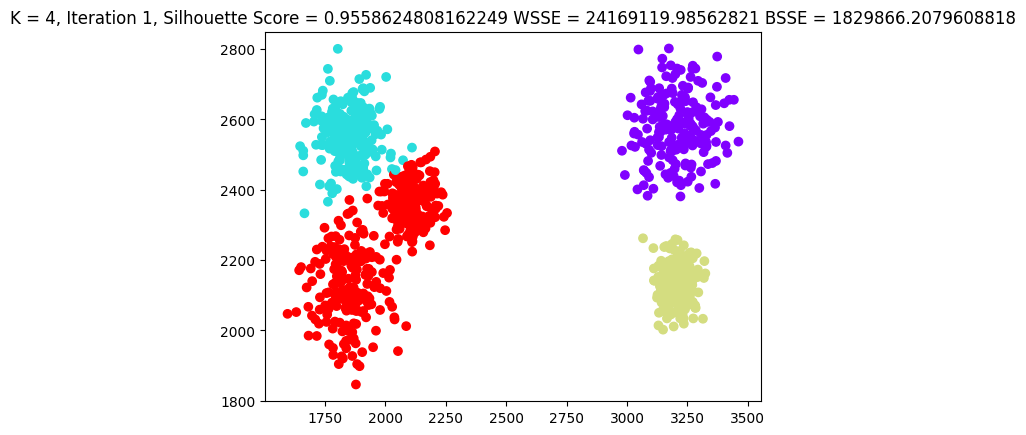
The detailed findings are given below:

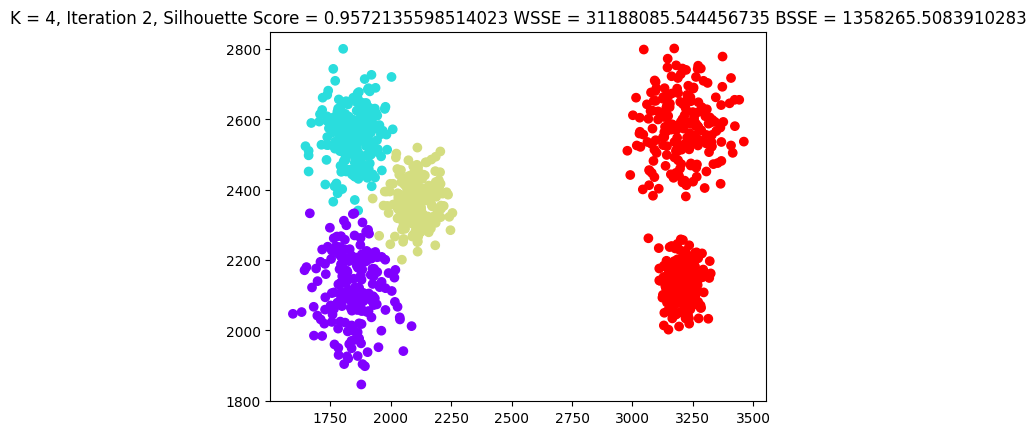


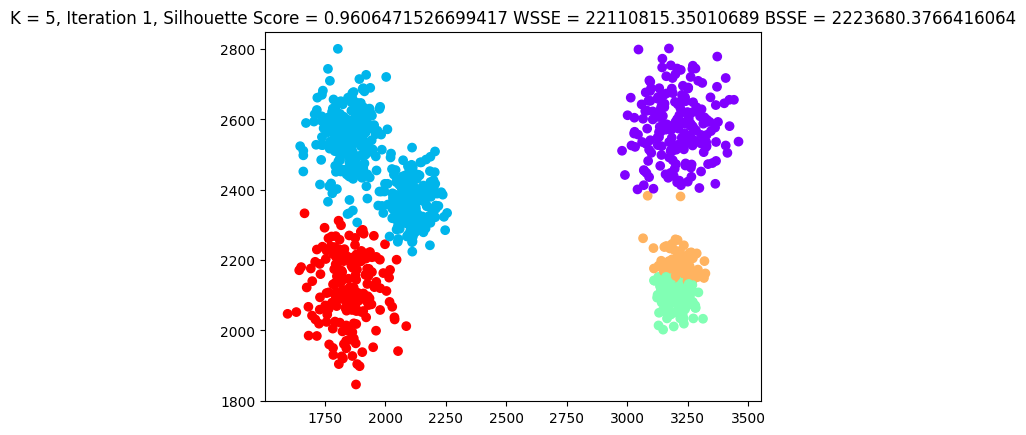


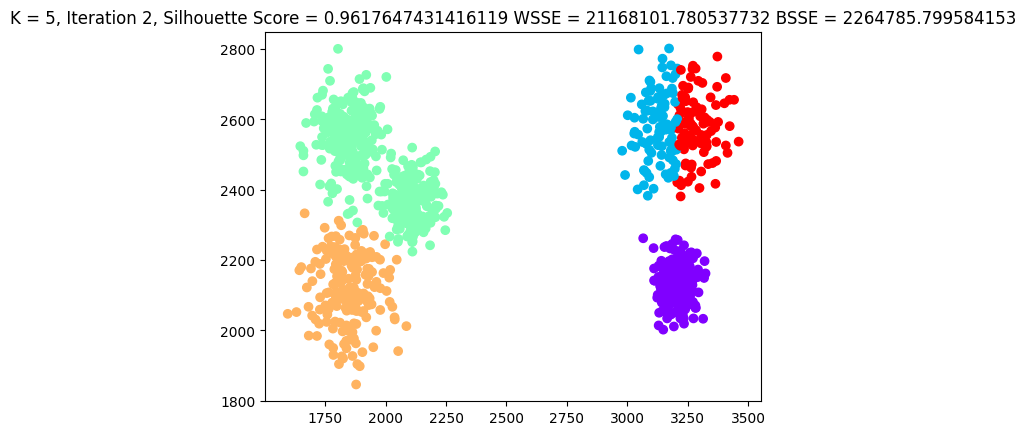


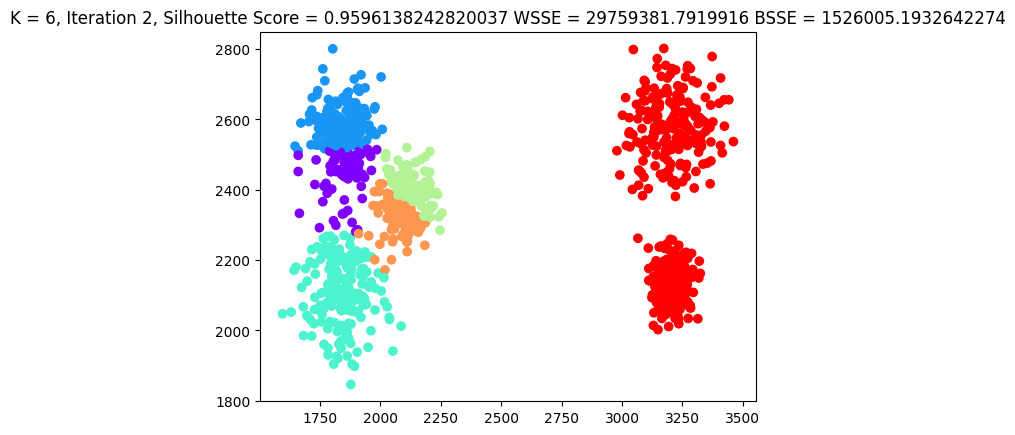
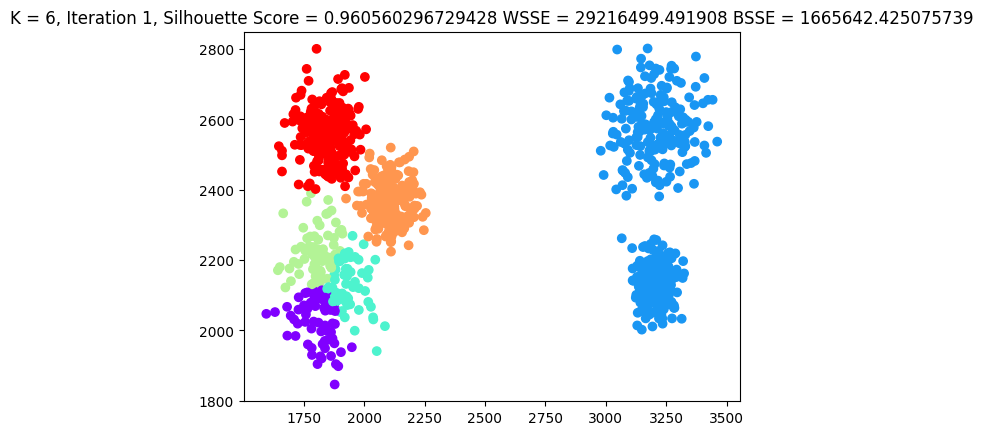


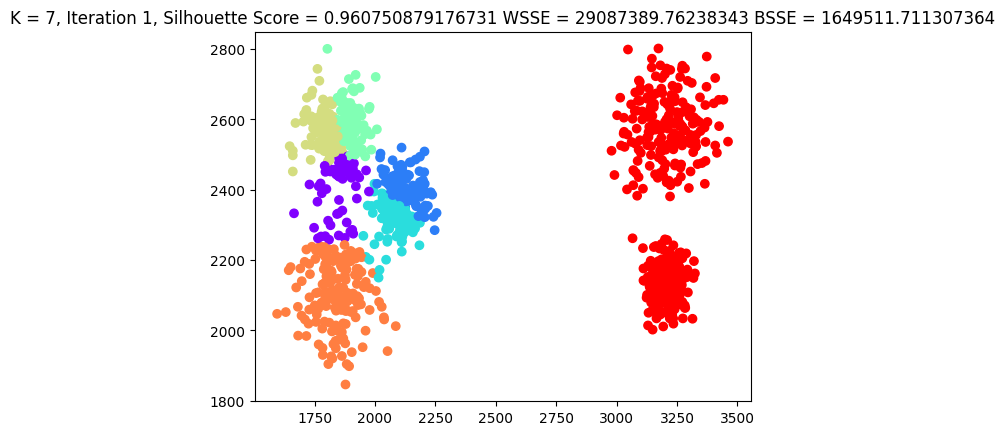


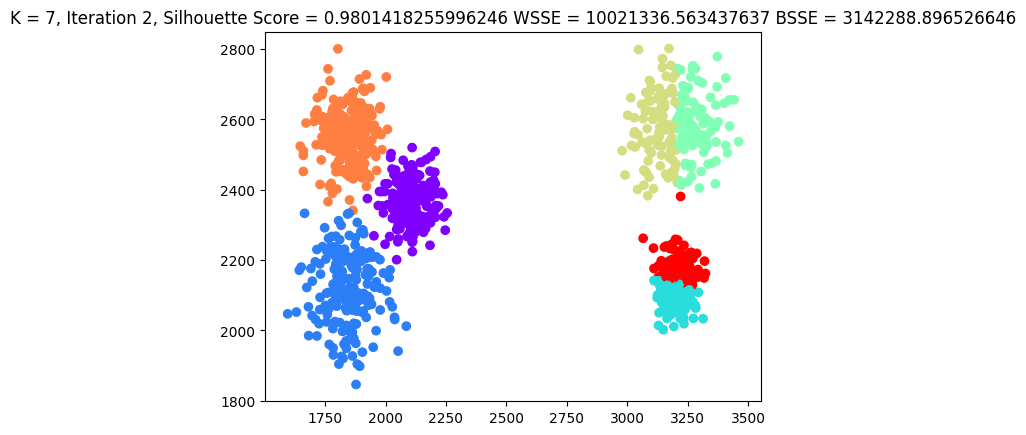


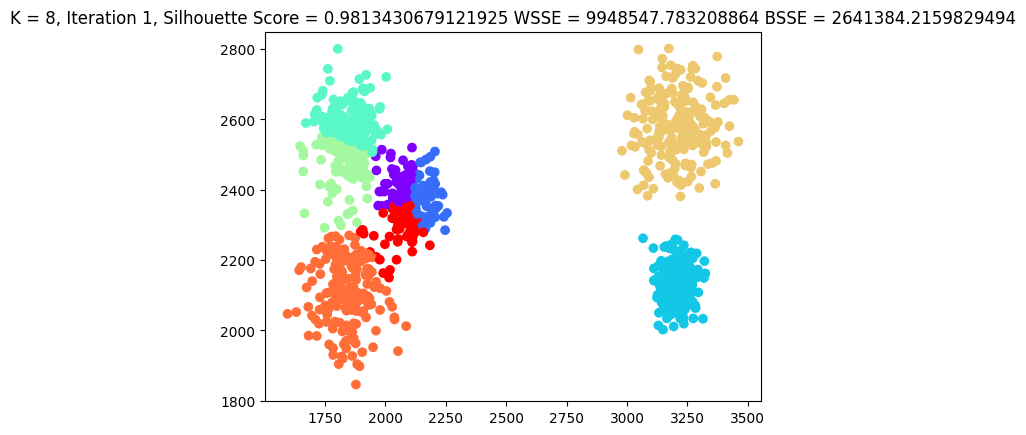


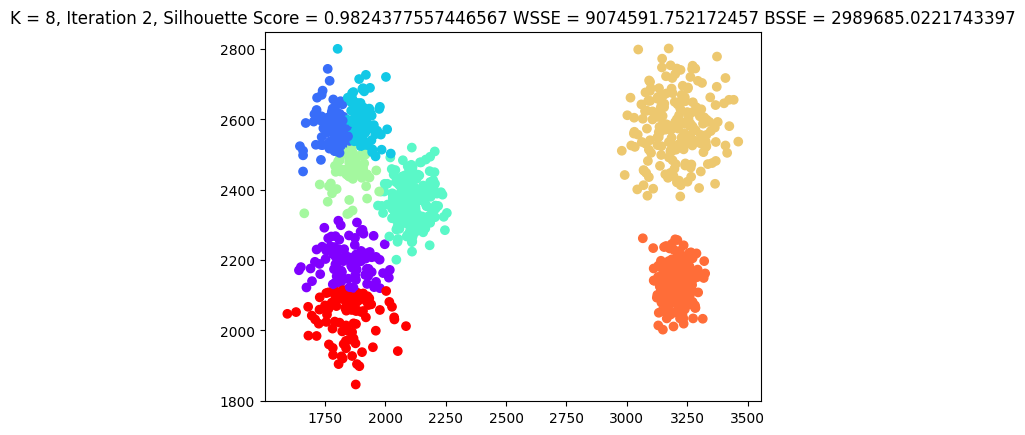


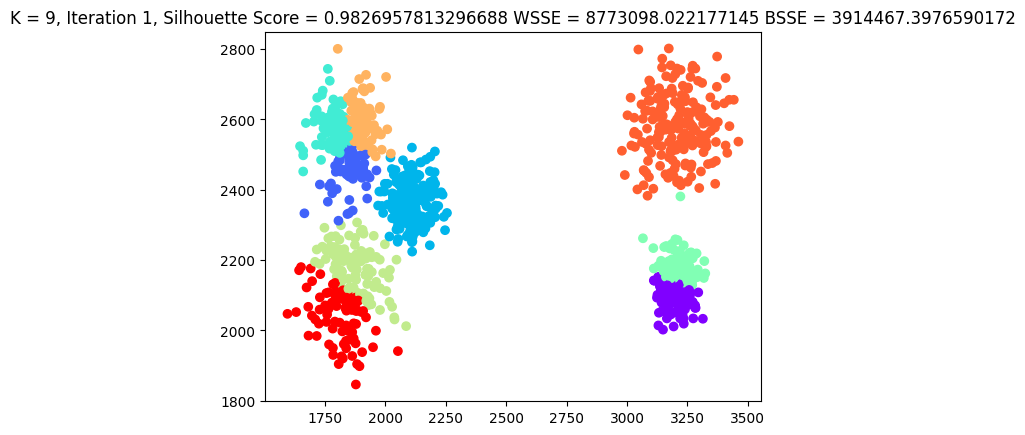


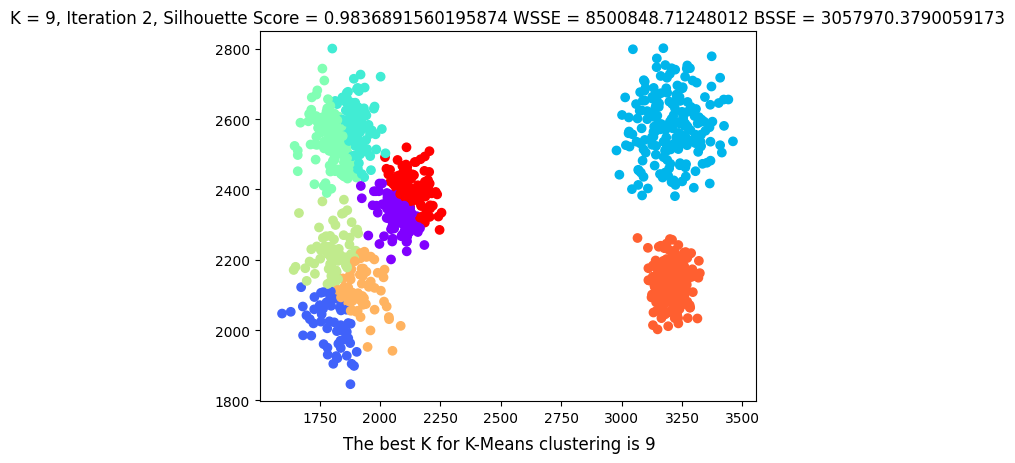






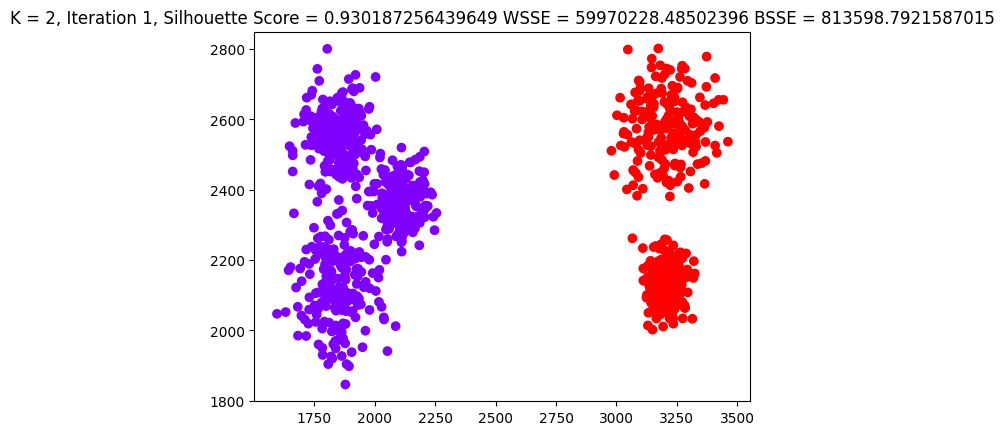


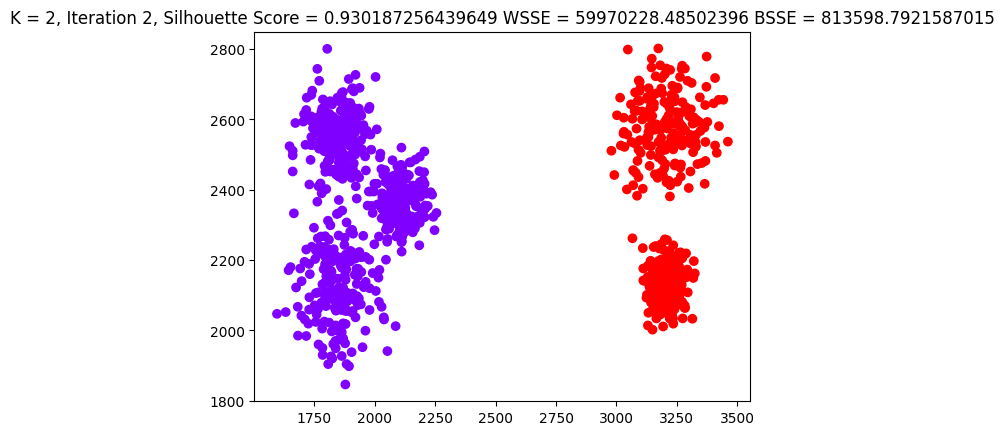


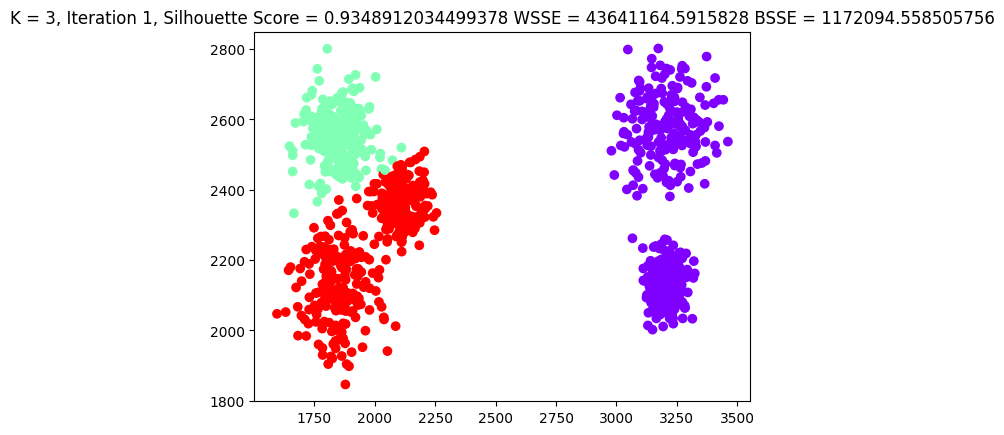


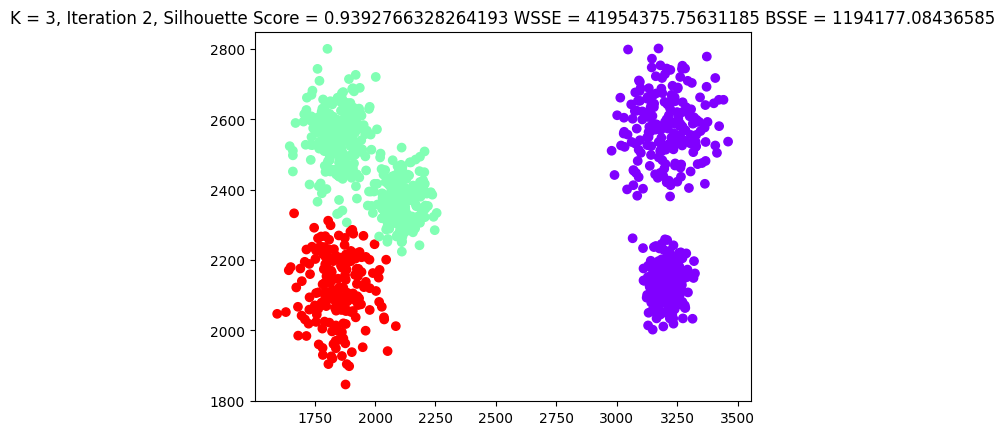
DS1.txt – Self-coded Bisecting Kmeans PYSPARK  
Running the self-coded Bisecting Kmeans clustering on DS1.txt yielded *K = 9* as the best number of clusters when clusters were tested by taking 2 to 9 centroids (quality metric = Silhouette Score*. K = 9* had the best score which was **0.98405406**). Bisecting Kmeans performed better clustering than which is shown by both the improved Silhouette scores and the plots.

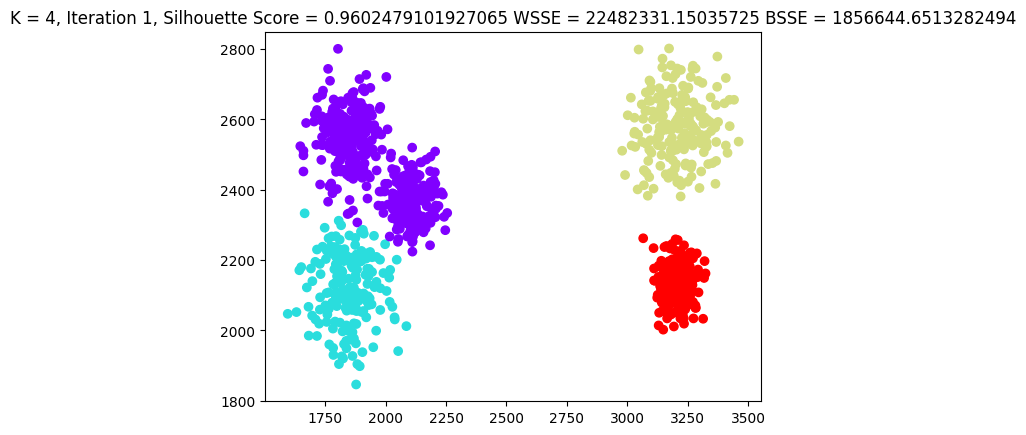
The detailed findings are given below:

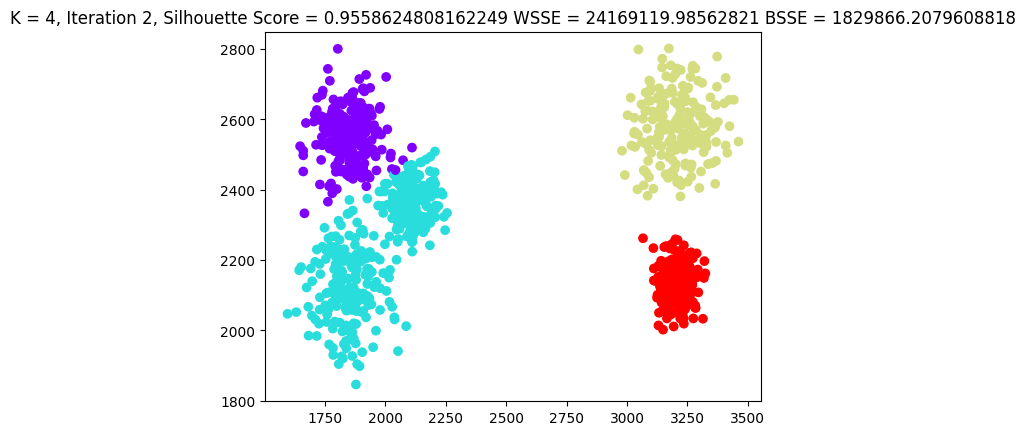


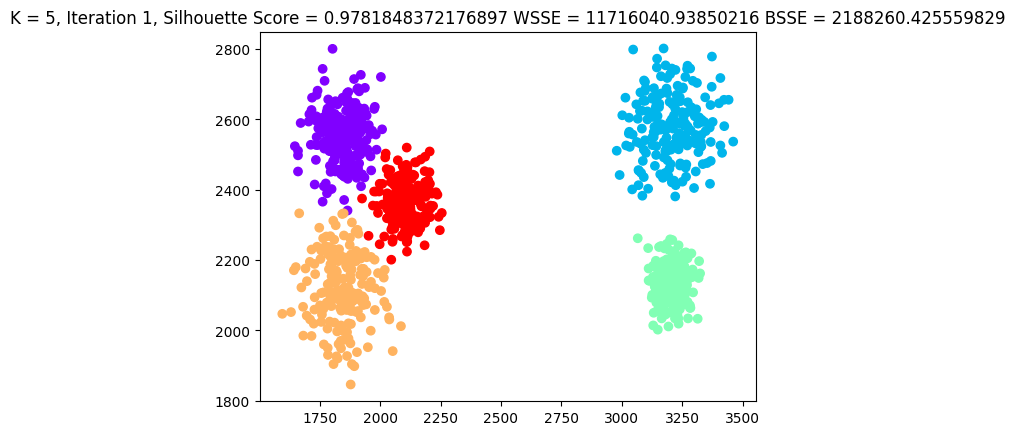


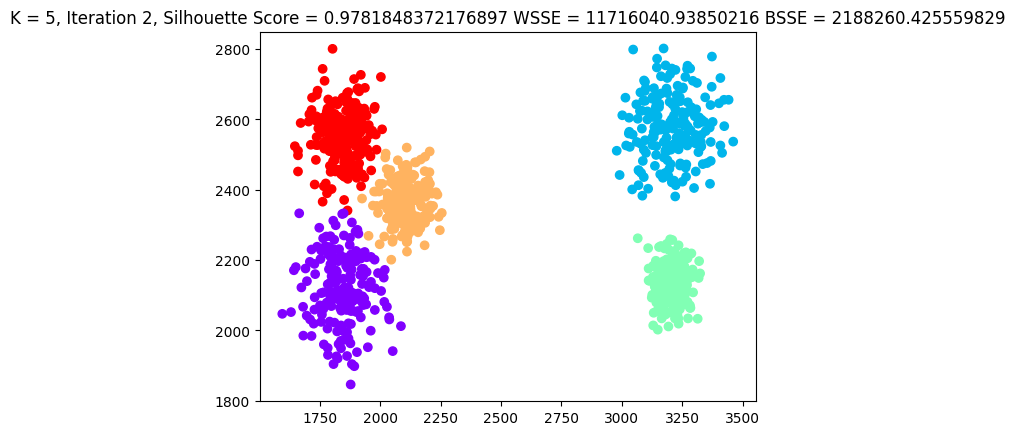


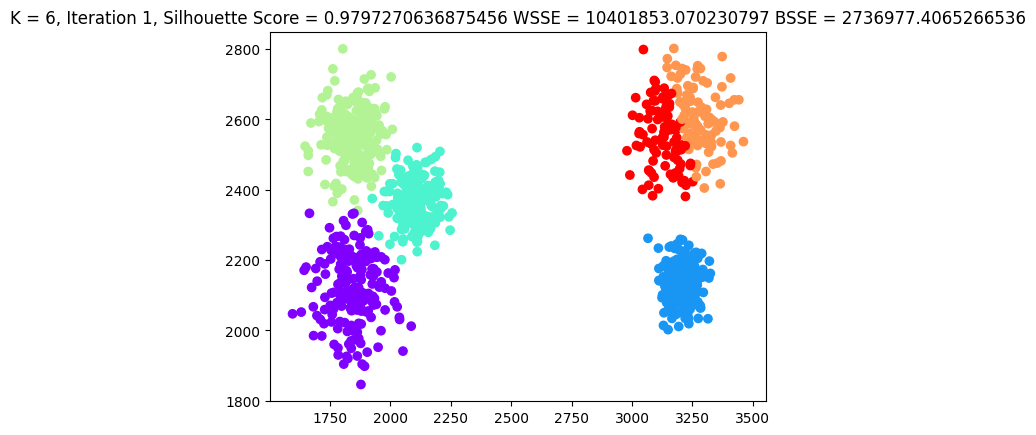


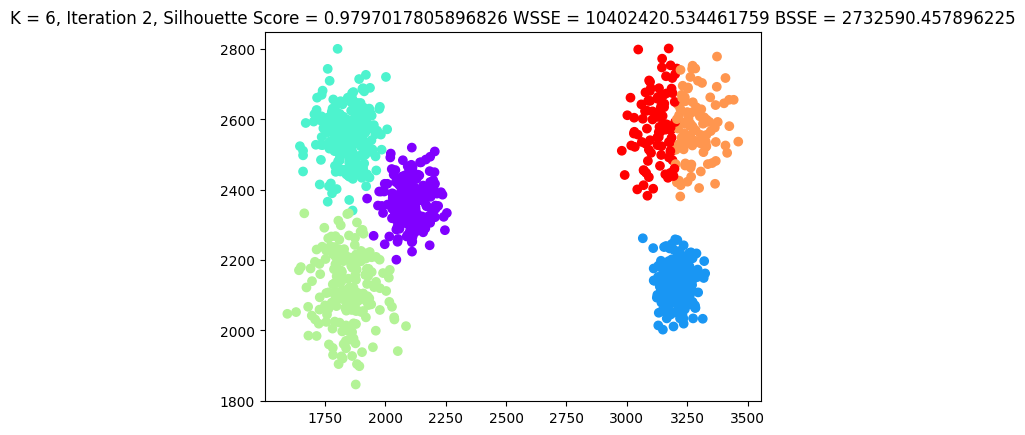


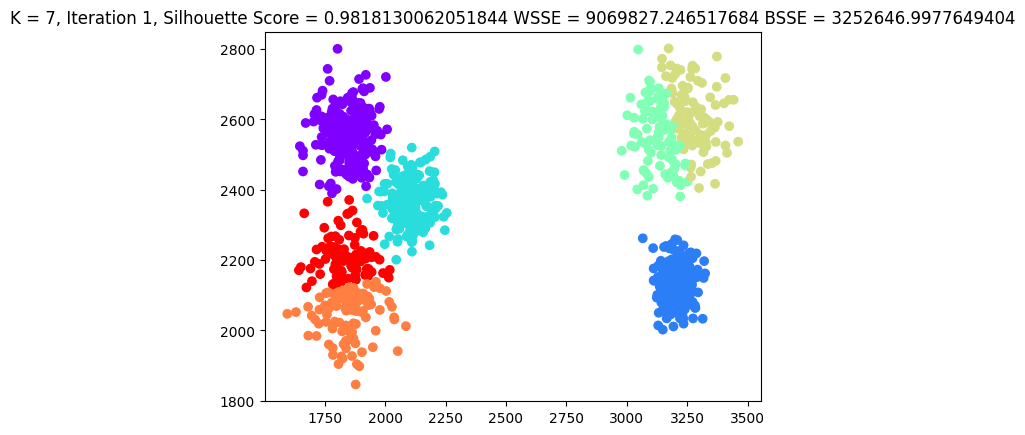


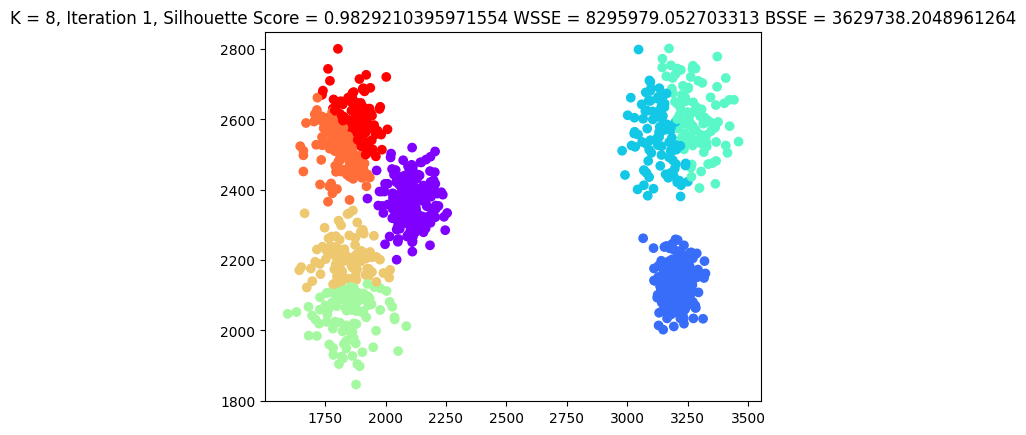


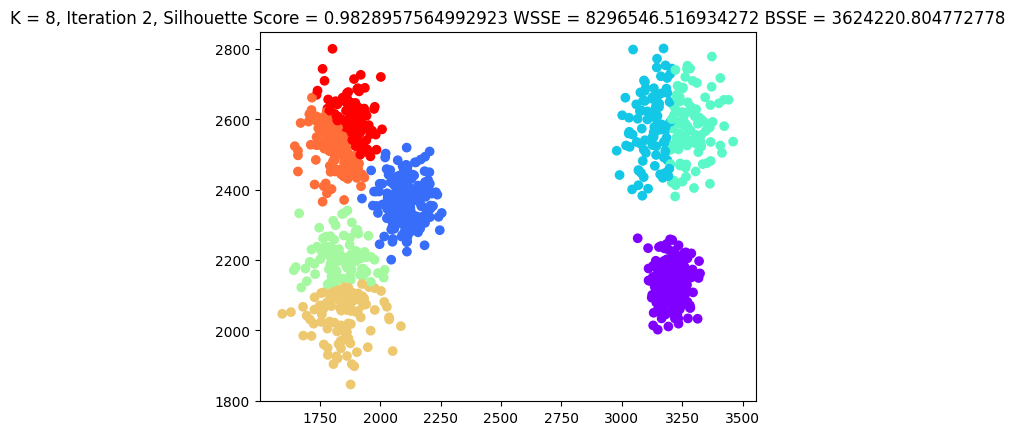


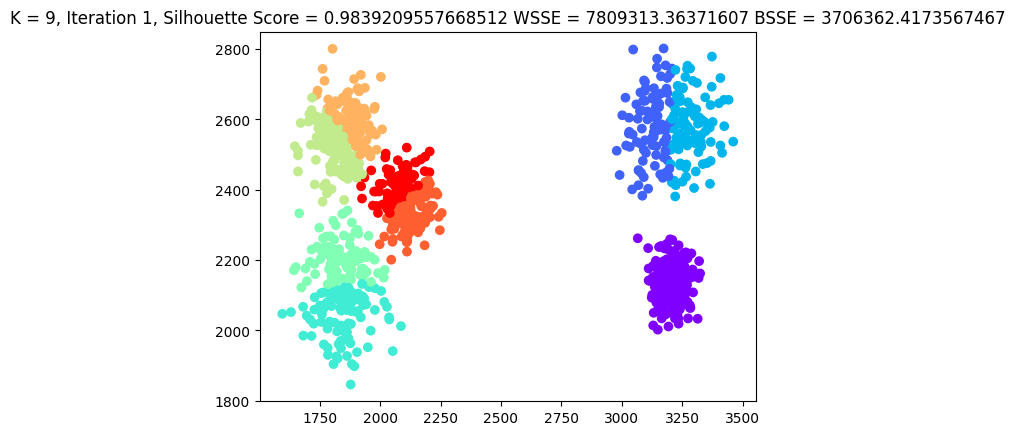


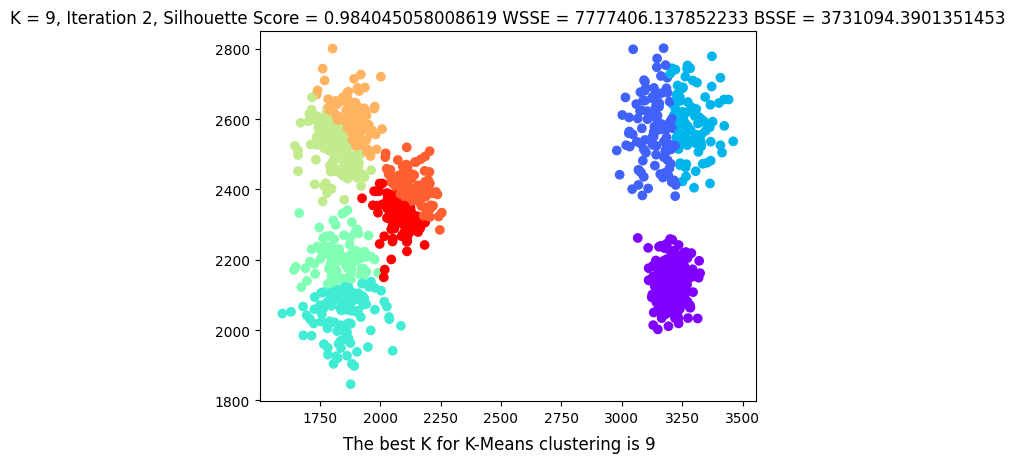






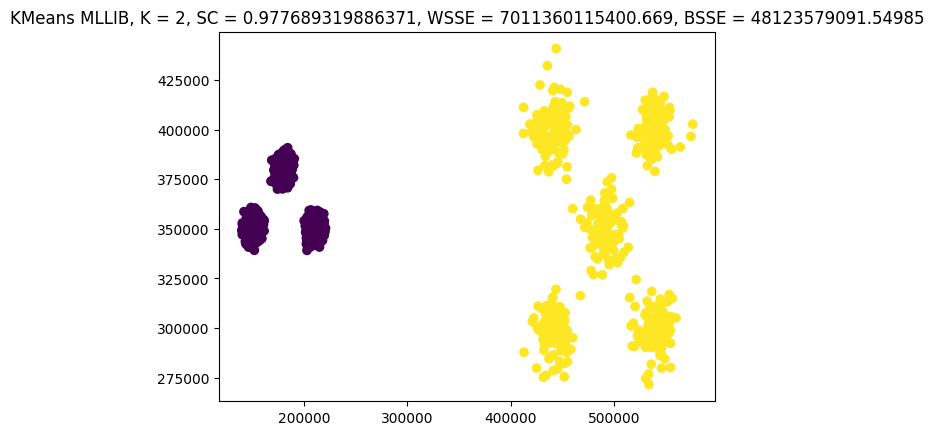


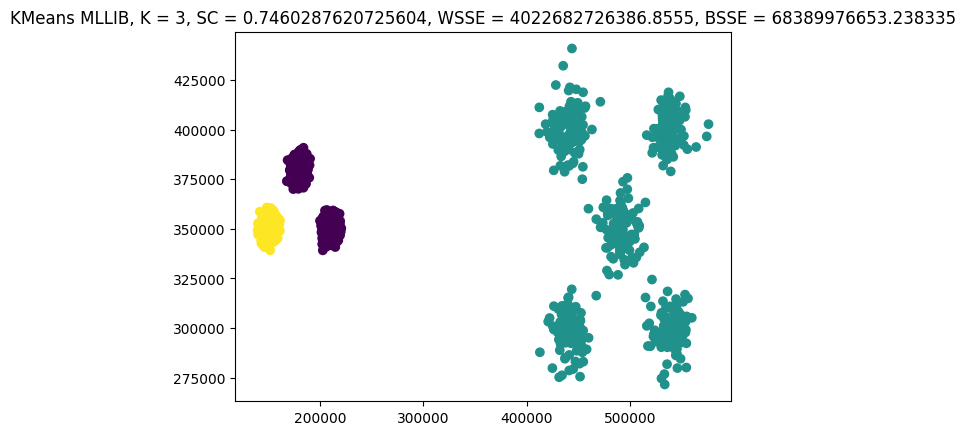


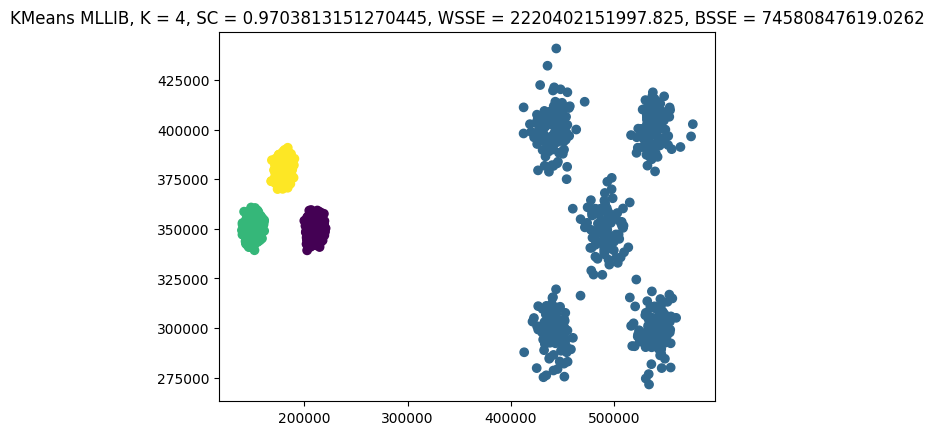


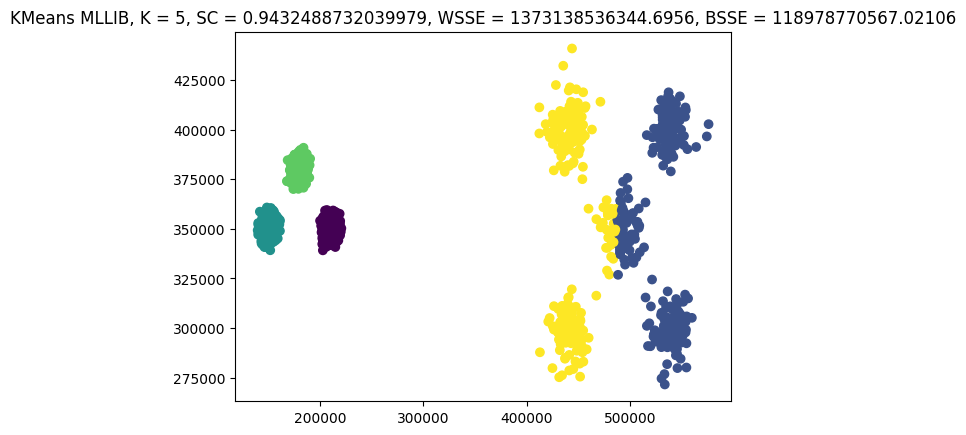
DS2.txt – Kmeans MLLIB library PYSPARK

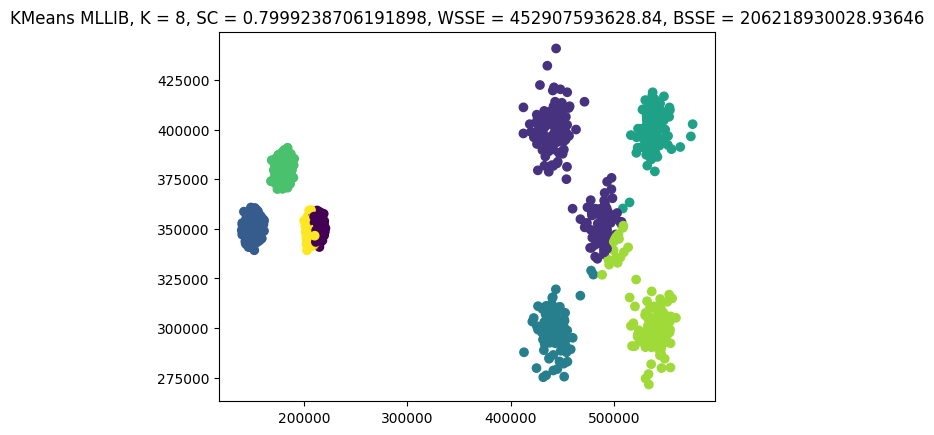
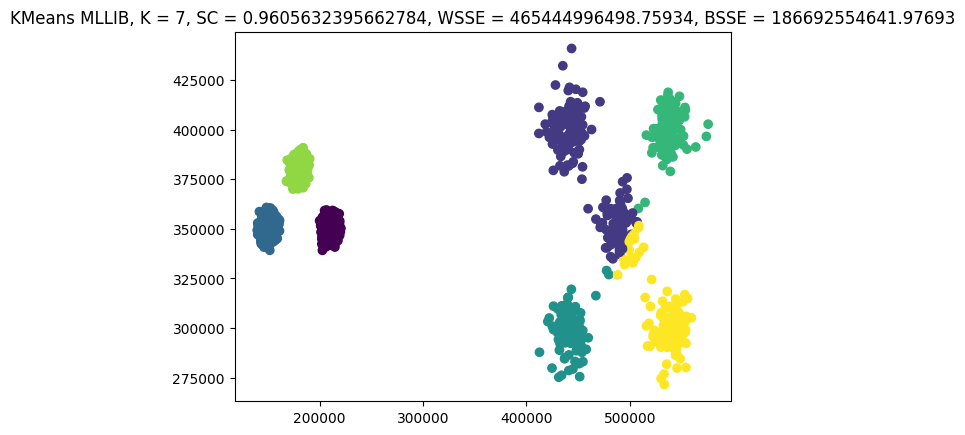
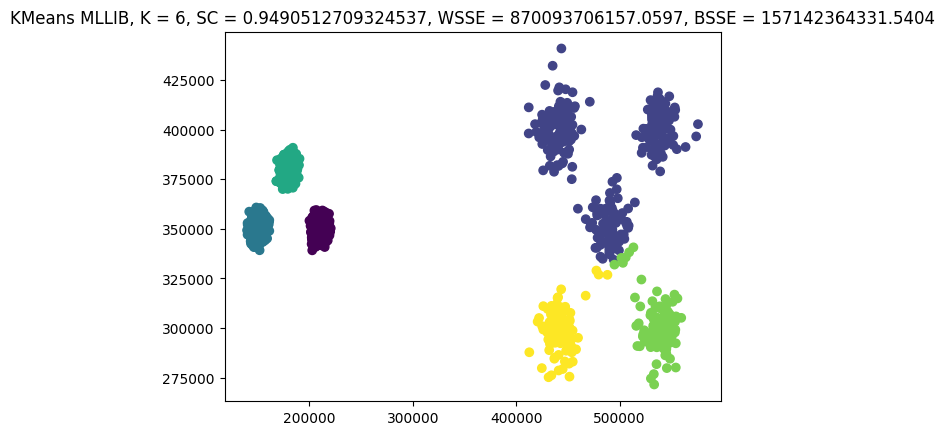
Performing Kmeans clustering using PYSPARK MLLIB on DS2.txt yielded *K = 2* as the best number of clusters when clusters were tested by taking 2 to 9 centroids (quality metric = Silhouette Score*. K = 2* had the best score which was **0.977689**).

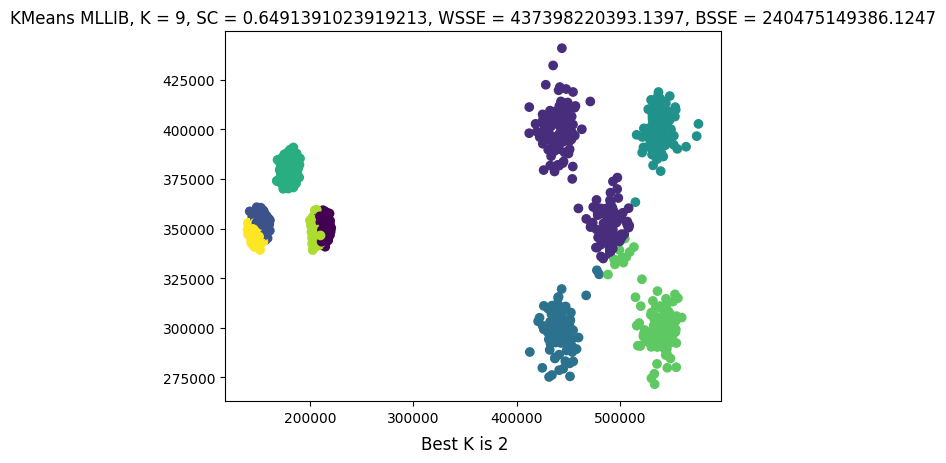






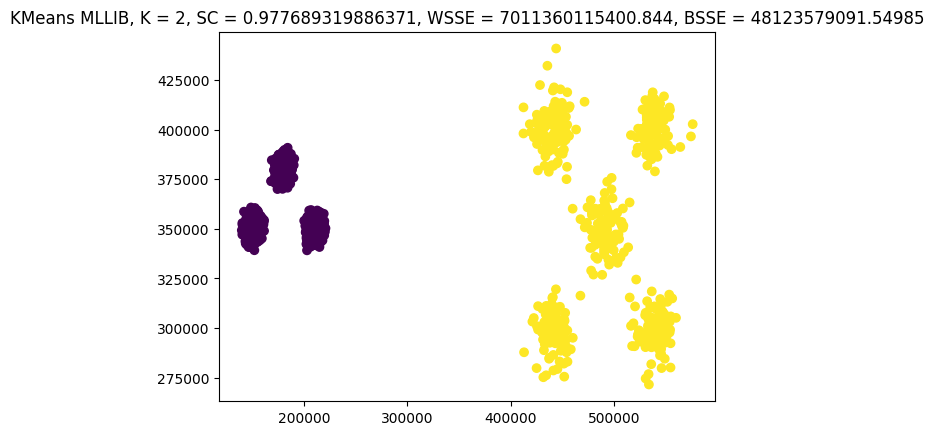


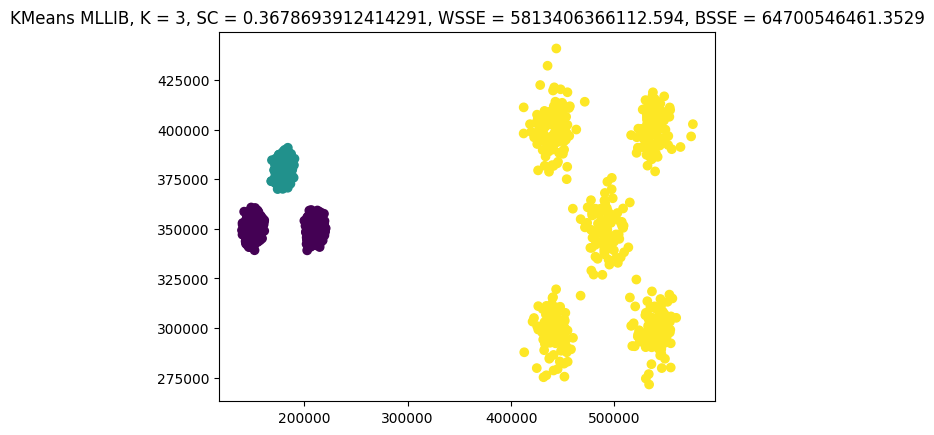


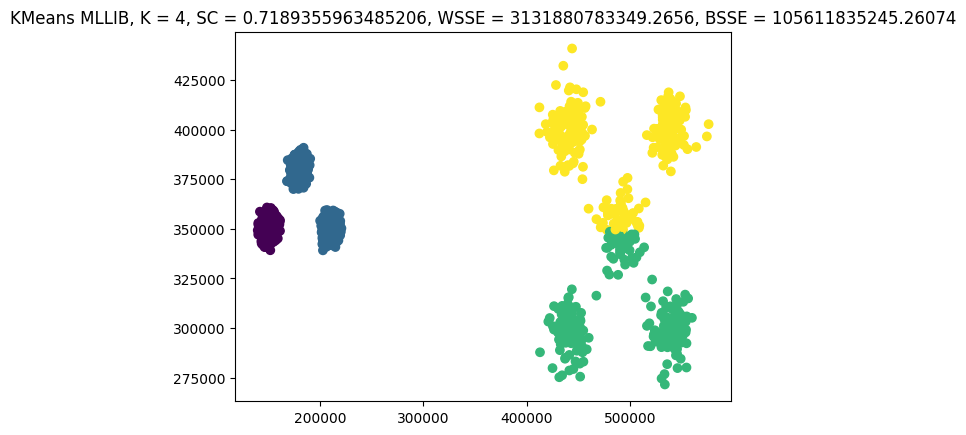


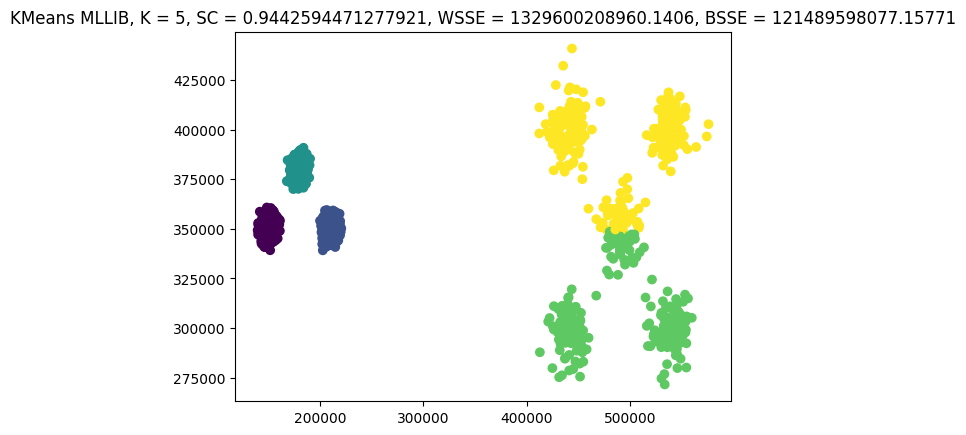
DS2.txt – Bisecting Kmeans MLLIB library PYSPARK

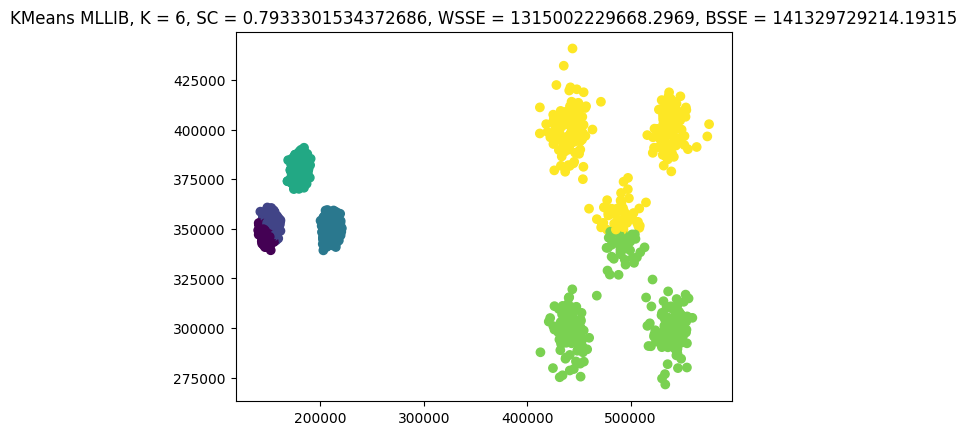
Performing Bisecting Kmeans clustering using PYSPARK MLLIB on DS2.txt yielded *K = 2* as the best number of clusters when clusters were tested by taking 2 to 9 centroids (quality metric = Silhouette Score*. K = 2* had the best score which was **0.977689**).

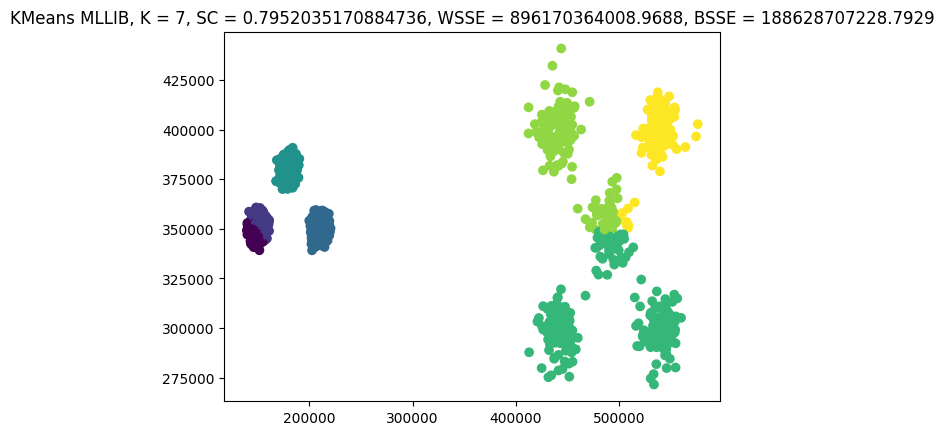


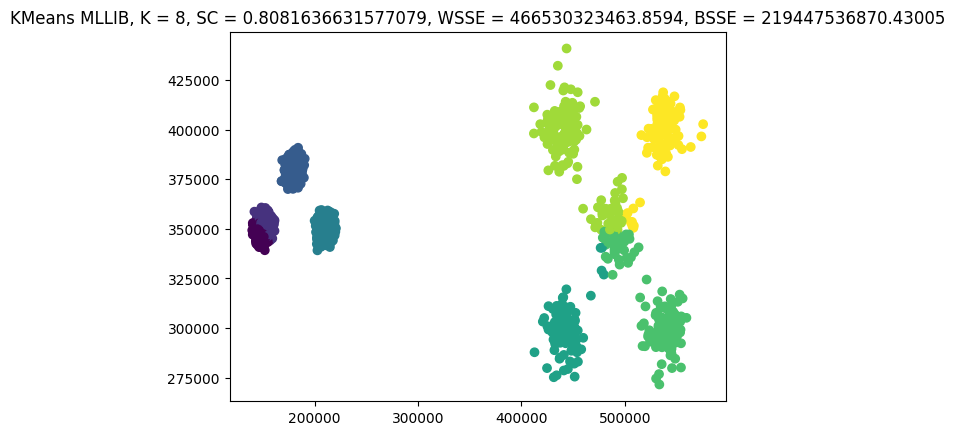


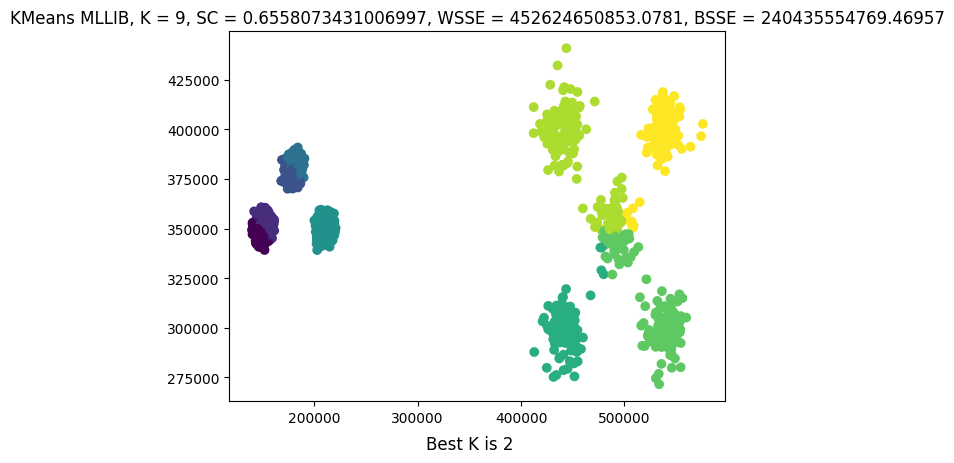






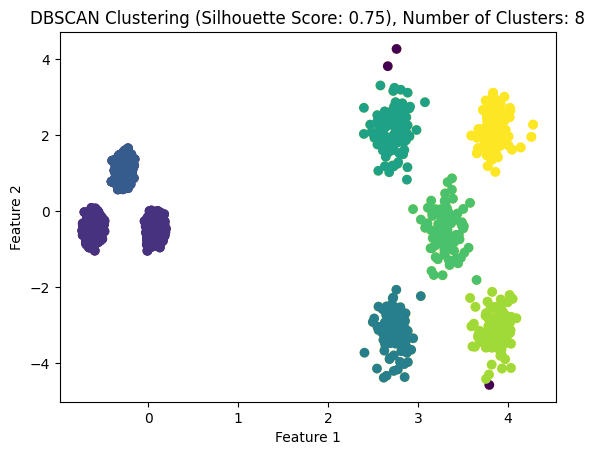






For DS2.txt, we can see that Kmeans performs better than Bisecting Kmeans until K=7 based on the Silhouette Score measure. However, as we increase the number of clusters, Bisecting Kmeans starts to perform better. For clusters 2 to 9, the best number of clusters given by Silhouette score for both clustering types is 2 and yields similar Silhouette Scores.

Post Processing – DBSCAN – Kmeans



Post Processing – DBSCAN – Bisecting Kmeans

