Note: Using the full dataset of 20000 observations, it takes >30 minutes to output calculateSC() when the parameters of cluster() are 50 for # of iterations and 32 # of clusters, respectively. When using the first 2000 observations, however, the runtime no more than 2 minutes per call.

SC Averages

k SC\_full SC\_4 SC\_2

32 0.4109 0.4198 0.4008

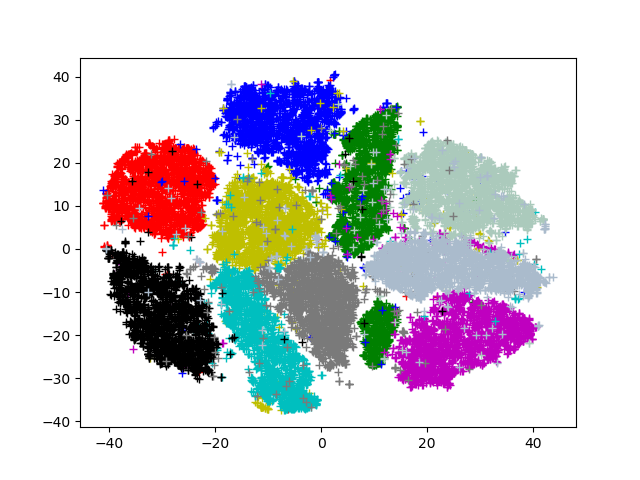
16 0.4058 0.4274 0.4184

8 0.4218 0.4653 0.4198

4 0.3948 0.696 0.6102

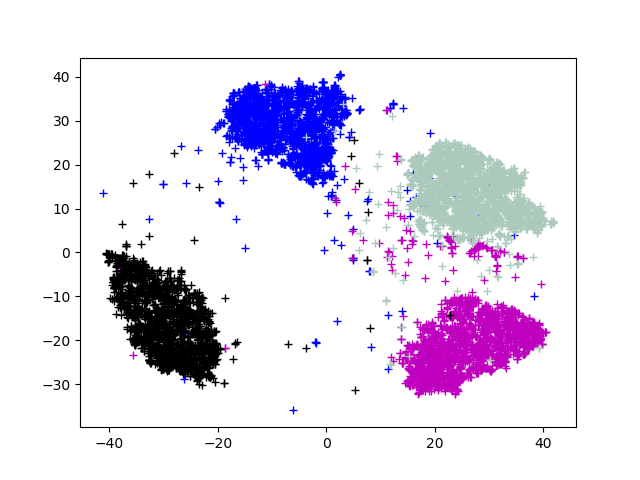
2 0.3738 0.4612 0.8315

Full Data Set



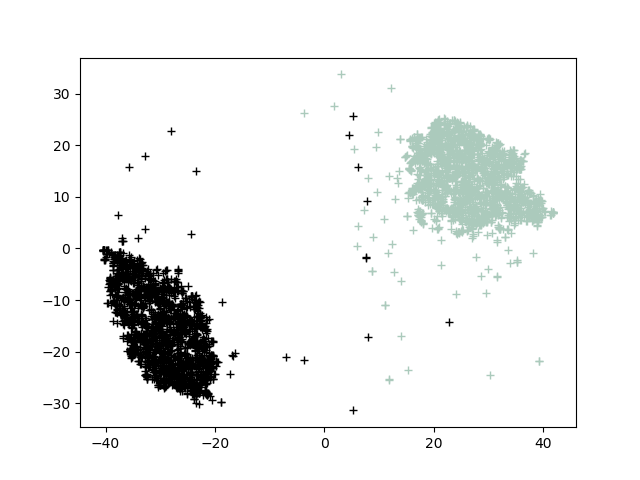
Choose k = 8, since 8 has the highest SC, and there are around 10 clusters.

Partial Data Set {2, 4, 6, 7}



Choose k = 4, since 4 has the highest SC and there are exactly 4 clusters.

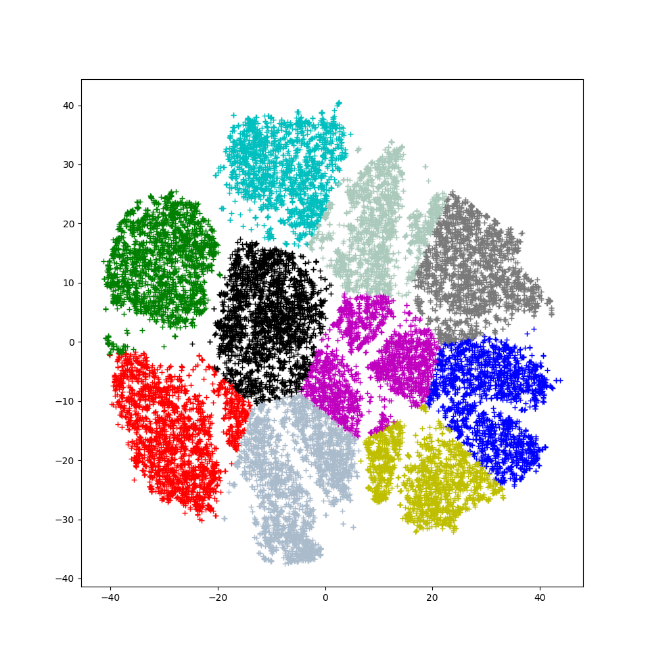
Partial Data Set {6, 7}



Choose k = 2, since 2 has the highest SC, and there are exactly 2 clusters.

Bonus results:

100 iterations when k = 10 [7 minute runtime]



In additional to using print statements, to test the validity of the calculateSC() function, we will use the first 100 observations as the data set.

At 50 iterations, calculateSC() approaches 1.0 when k approaches 100

At k = 2, SC = .394

At k = 5, SC = .439

At k = 10, SC = .495

At k = 50, SC = .673

At k = 90, SC = .919

At k = 95, SC = .972

At k = 100, SC = 1.0