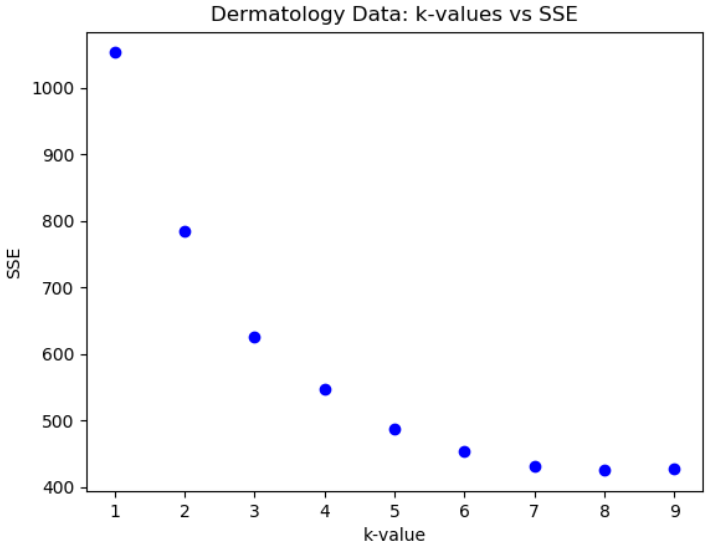
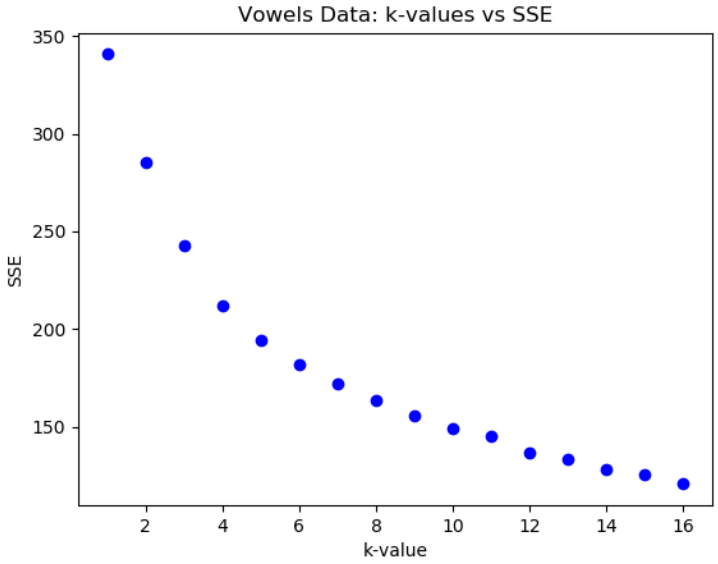
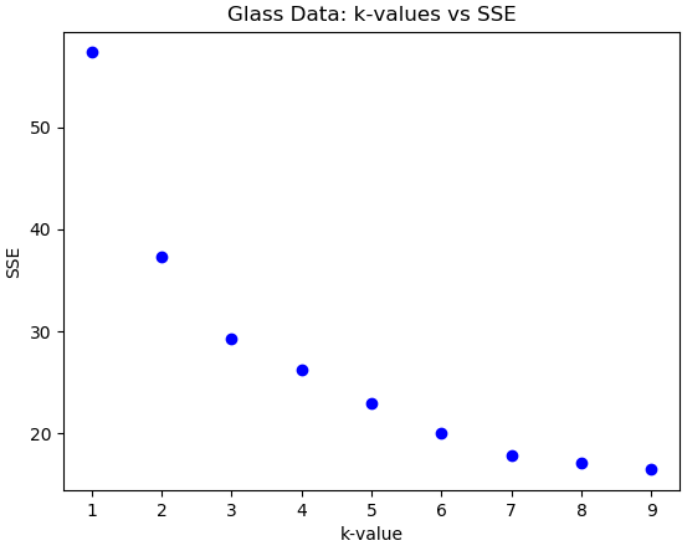
Assignment 5 – Clustering

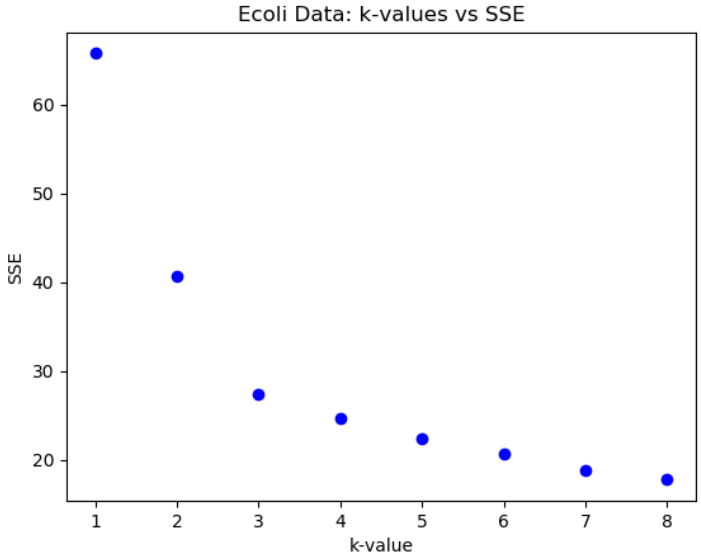
3 – K-Means Clustering

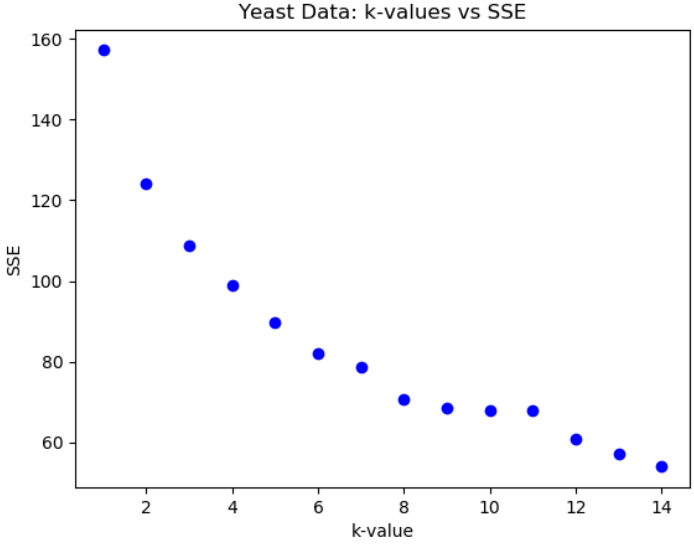
3.2.1 – SSE vs K for each Dataset:

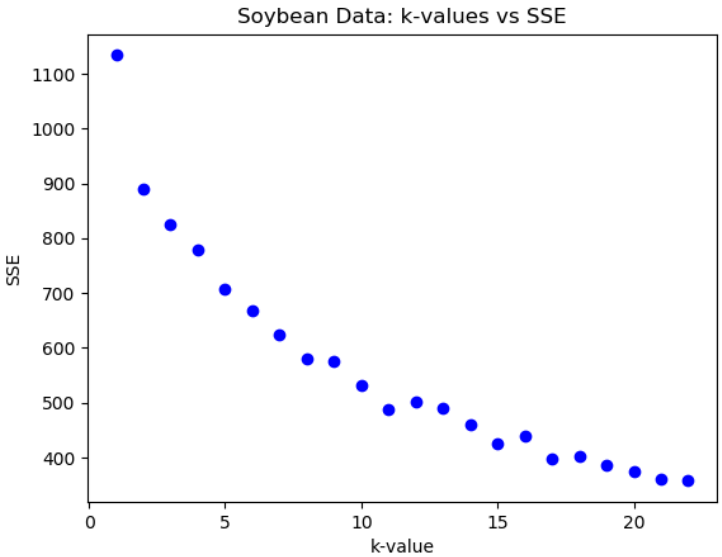












3.1.3: Optimal number of clusters is chosen based on the “elbow”-method, when SSE does not decrease as much when k-value increases.

3.2.2 – Optimal Number of Clusters Based on SSE and corresponding NMI:

Dataset	Optimal Clusters based on SSE	Corresponding NMI
Dermatology	5	0.822
Vowels	5	0.298
Glass	6	0.311
E. coli	3	0.656
Yeast	9	0.276
Soybean	11	0.634

3.2.3 – Clusters = Number of Classes:

SSE and NMI values calculated as the average of 5-folds starting with random initial means.

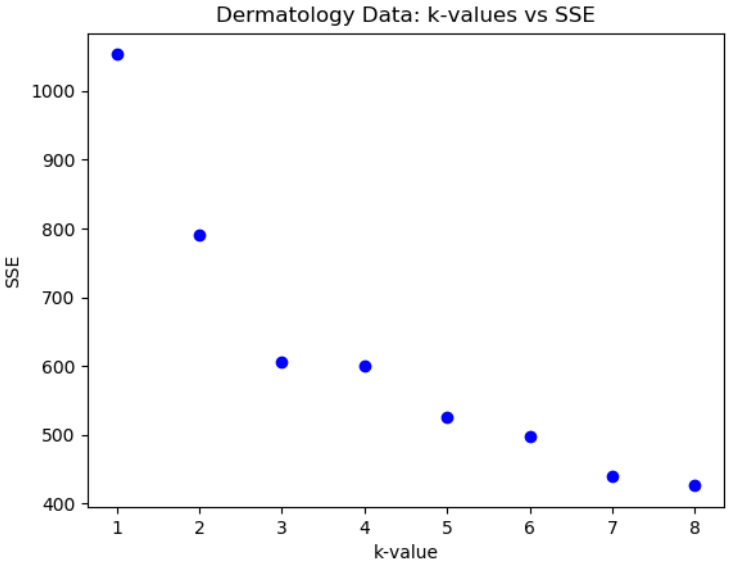
Dataset	Classes/ Clusters	SSE	NMI
Dermatology	6	454.04	0.825
Vowels	11	145.37	0.368
Glass	6	19.97	0.311
E. coli	5	22.31	0.614
Yeast	9	68.47	0.276
Soybean	15	425.00	0.672

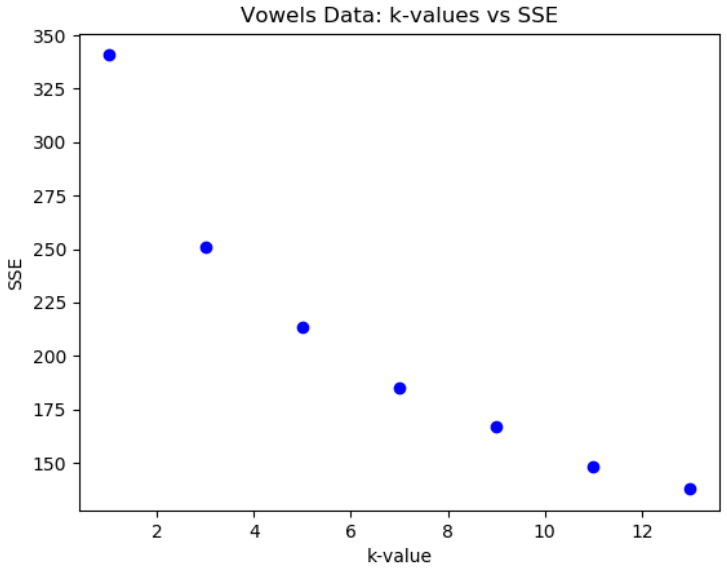
4 – Gaussian Mixture Models (GMM)

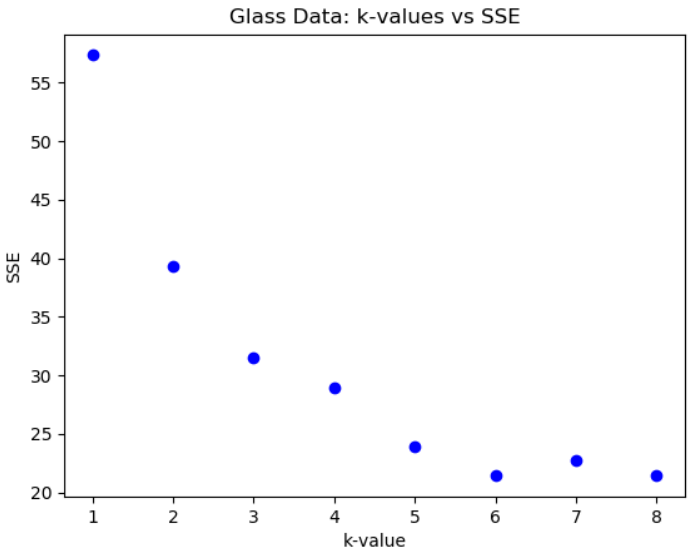
4.1.4 – SSE vs NMI for GMM?

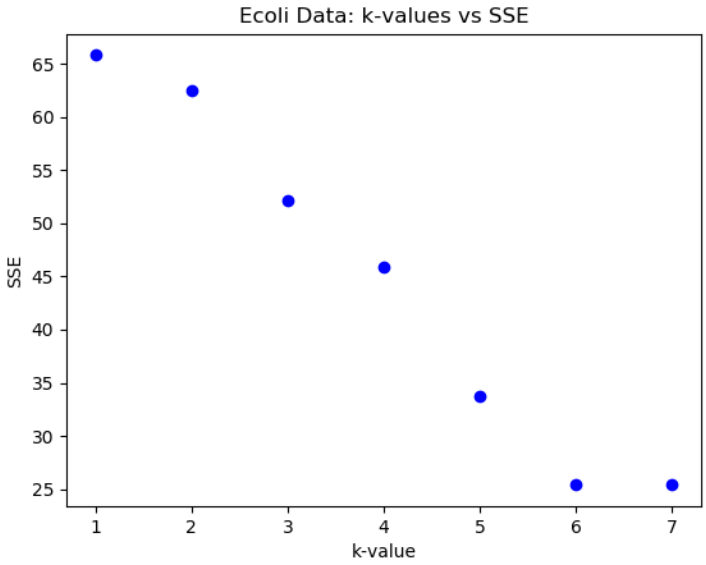
NMI is a better criterion for GMM than SSE, because GMM uses Mean as well as Variance to determine the best cluster for a data-point to belong to, and the SSE only takes into account the mean value. NMI value, rather than relying on mean or variance, is based on cluster purity compared to given class labels, and is a more effective measure of how effective clustering using GMM is than if using SSE.

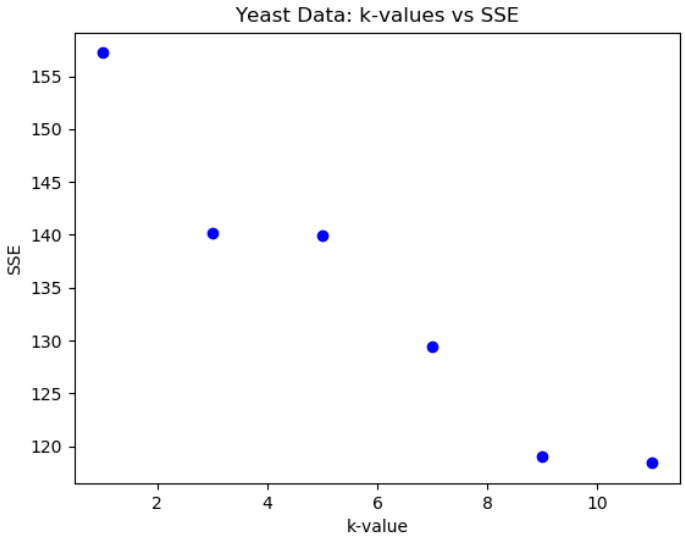
4.2.1 – SSE vs k plots:

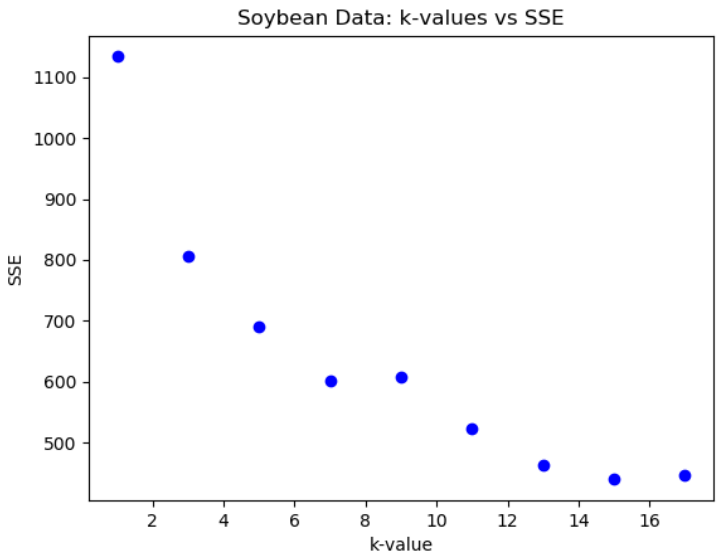




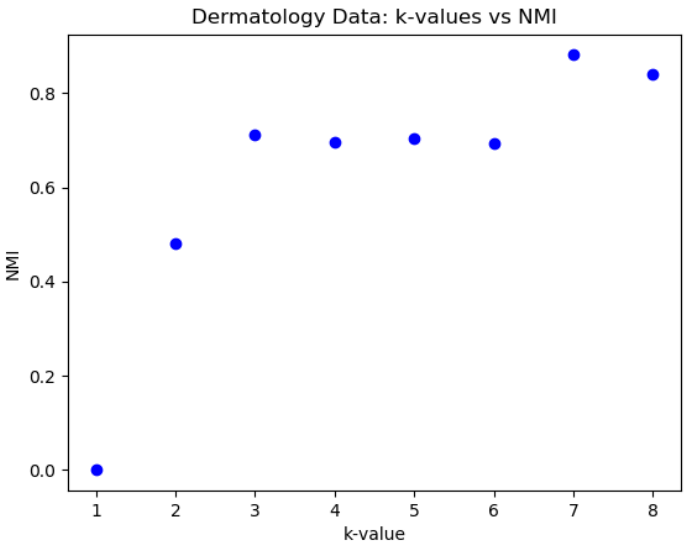


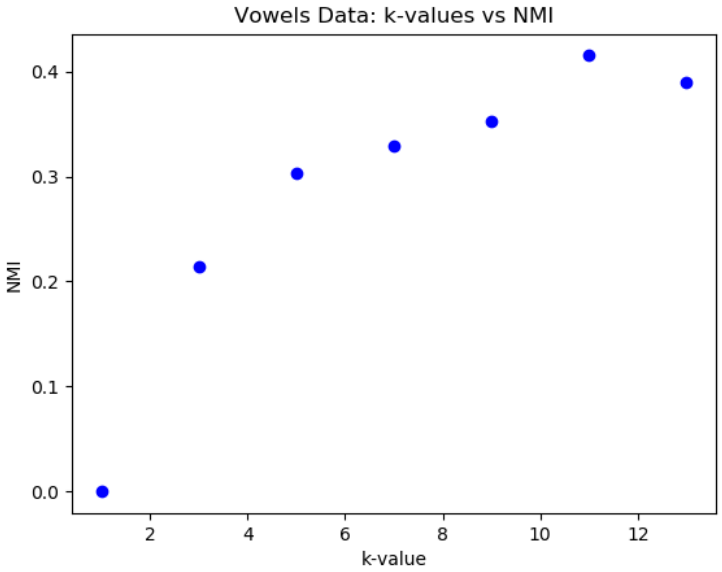


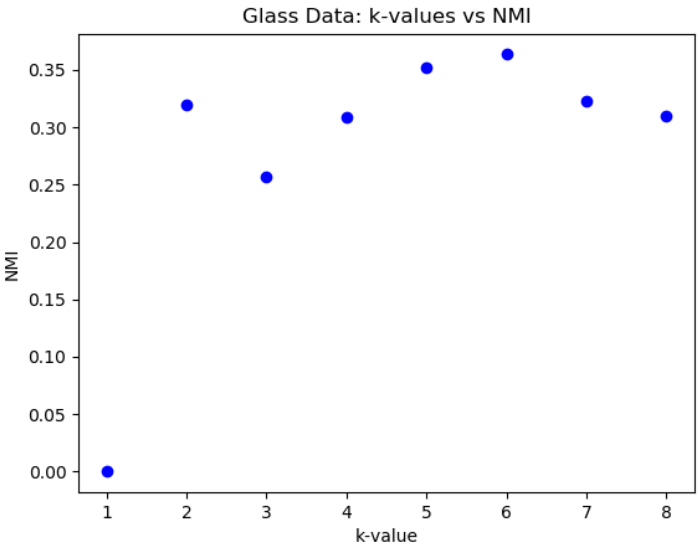


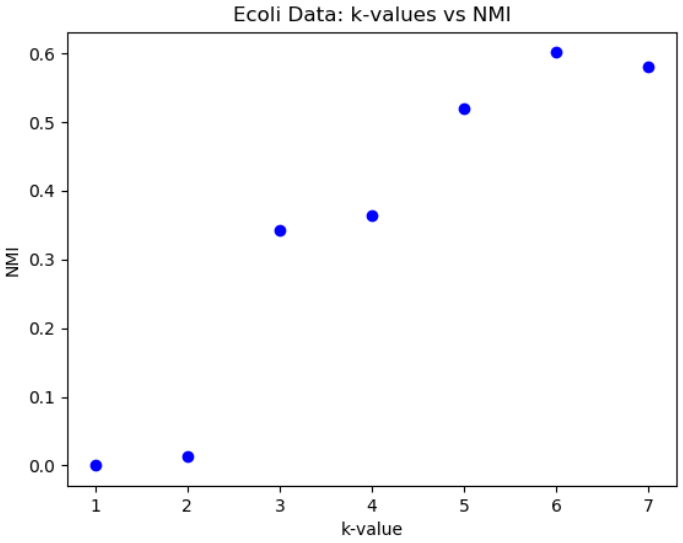


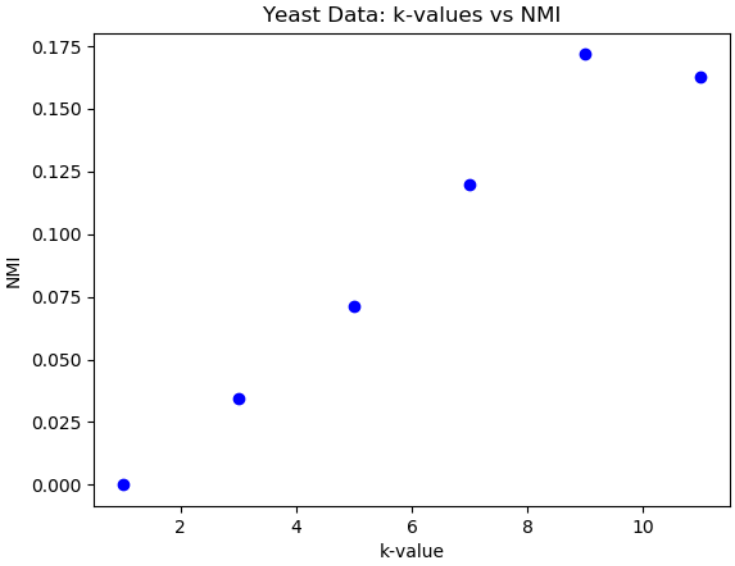
4.2.2 – NMI vs k plots:

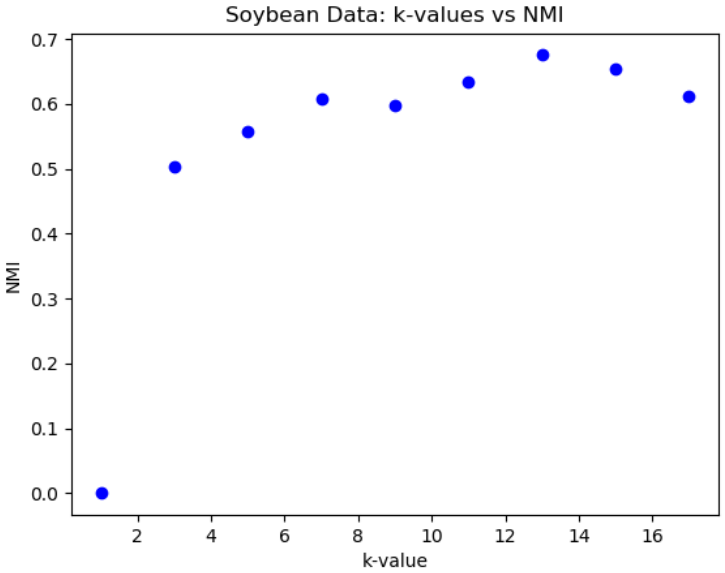












4.2.3 – Optimal Clusters and NMI based on SSE:

Dataset	Optimal Clusters based on SSE	Corresponding NMI
Dermatology	5	0.705
Vowels	11	0.415
Glass	6	0.363
E. coli	6	0.601
Yeast	9	0.172
Soybean	13	0.675

4.2.4 – Optimal Clusters and SSE based on NMI:

Dataset	Optimal Clusters based on NMI	Corresponding SSE
Dermatology	7	440.81
Vowels	11	148.13
Glass	6	21.40
E. coli	6	25.40
Yeast	9	119.01
Soybean	13	463.05

4.2.5 – Clusters = Classes and Corresponding NMI:

Dataset	Classes/Clusters	NMI
Dermatology	6	0.695
Vowels	11	0.415
Glass	6	0.363
E. coli	5	0.519
Yeast	9	0.172
Soybean	15	0.654

5 – Comparing k-Means and GMM

5.1 – k-Means vs GMM per Dataset:

Dataset	Preferred Algorithm
Dermatology	k-Means – Because there is little difference in performance between k-Means and GMM for this dataset, so clustering with k-Means with uniform variance (circular clusters) seems to result in similar clusters with GMM even when using variance.
Vowels	GMM – k-Means had poor performance, so taking into account variance and normal distribution should increase NMI score and clustering performance.
Glass	GMM – GMM was only a slight improvement over k-means here, but it did increase NMI scores across the board at each k-value.
E. coli	k-Means – E. coli gave decent results of NMI scores using k-Means value with almost no improvement when using GMM algorithm.
Yeast	GMM – When using GMM, the NMI very clearly peaked at the same k-value as the number of classes in the original data. This was not as clearly defined when using just k-Means.
Soybean	k-Means – Had almost identical performance when using GMM vs k-means so that indicates to me that k-Means algorithm is sufficient for this dataset, and that GMM is unnecessary.

5.2 – Insight into separability via Clustering:

Dataset	Insight into separability
Dermatology	Because k-means was very effective for this dataset (after min-max normalization) we can say that it's probable that there is little overlap between classes in this dataset, at least among some distinct features.
Vowels	The NMI peaks here during GMM at k=11 which indicates that the 11 unique classes are all separable, although the clusters generated were only partially accurate nor resulting in pure clusters.
Glass	For k-means and GMM the SSE and NMI values very clearly indicated that the most effective clustering happened when k=6. 6 is also the number of classes and that indicates that the classes are separable and that there shouldn't be much overlap between any of the 6 classes.
E. coli	The best k-values based on NMI and SSE was 6 (for GMM and k-means), while there were only 5 actual classes. This indicates to me that one of the classes probably could has two potential spreads within the class itself.
Yeast	The classes here are less separable than ideal, resulting in a low NMI for both GMM and k-Means, but better performance using GMM, because of taking into account variance. This indicates to me that there is significant overlap in the classes.
Soybean	Similarly to dermatology dataset, because k-means was very effective for this dataset we can say that it's probable that there is little overlap between classes in this dataset, at least among some distinct features.