Kolby Sarson
104232327

# Assignment 3

Running SciKit-Learn

Much the same as assignment 2, on SciKit-Learn I first imported 'pandas' to deal with the .csv files, 'numpy' and 'pyplot' to utilize and plot the data, and the clustering algorithm 'KMeans'. New to this assignment, I imported 'StandardScaler', 'PCA', and 'KernelPCA' for our feature extraction and dimensionality reduction. I wrote a function for each of PCA and KernelPCA to be applied to the datasets. I then loop through the datasets one by one, assigning a k-function and g-value (to optimize clustering), then apply the PCA/KPCA functions, reduce to 1-dimension, and cluster the 1D reduction result.
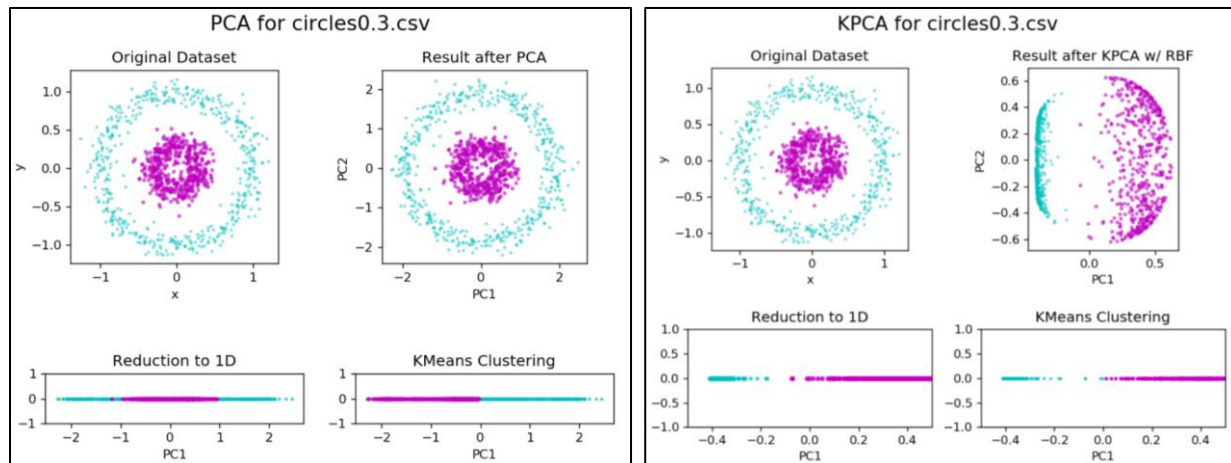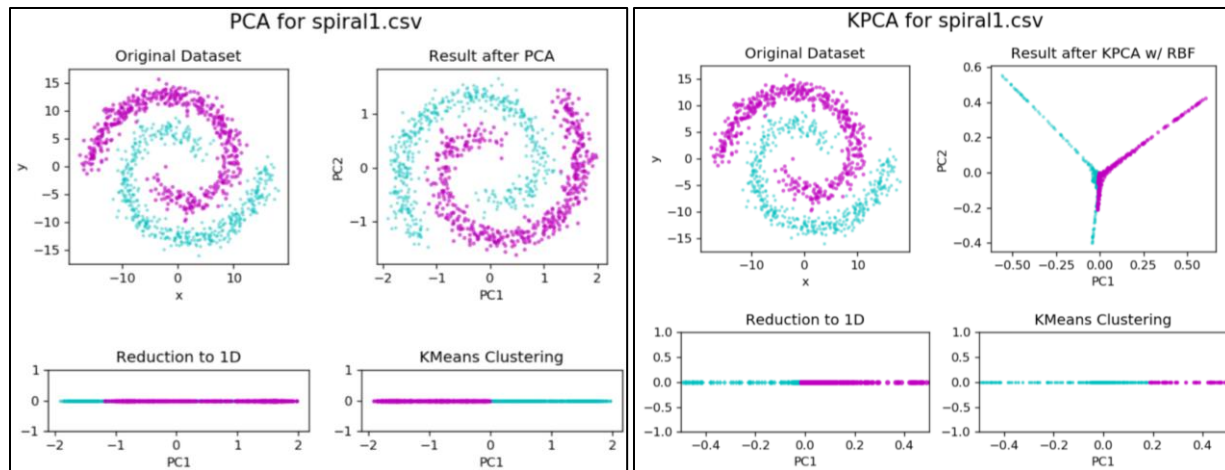

For the circles dataset, a gamma value of 1 was chosen. After applying dimensionality reduction with PCA, we see that the accuracy for the PCA Clustering was poor, visible by the rather different Reduction and Clustering plots (left bottom row). On the other hand, we can see that the accuracy for the KPCA Clustering was relatively high, as seen by the similarity with only slight differences between the KPCA's Reduction and Clustering plots (right bottom row).
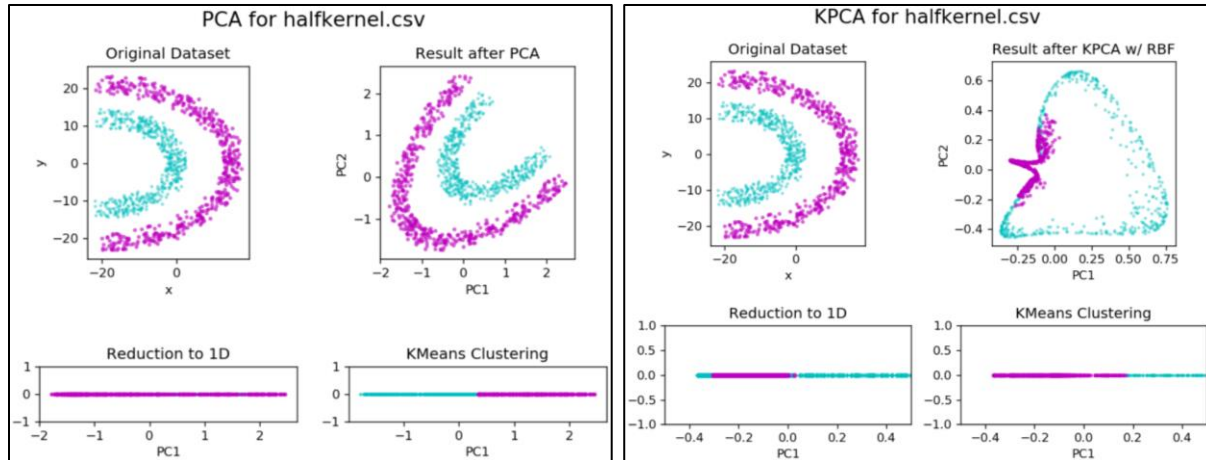
For the halfkernel dataset, a gamma value of 3 was chosen. After applying dimensionality reduction with PCA, we see again that the accuracy for the PCA Clustering was poor, visible by the rather different Reduction and Clustering plots (left bottom row). On the other hand, we can see that the accuracy for the KPCA Clustering was relatively high, as seen by the similarity with only slight differences between the KPCA's Reduction and Clustering plots (right bottom row).
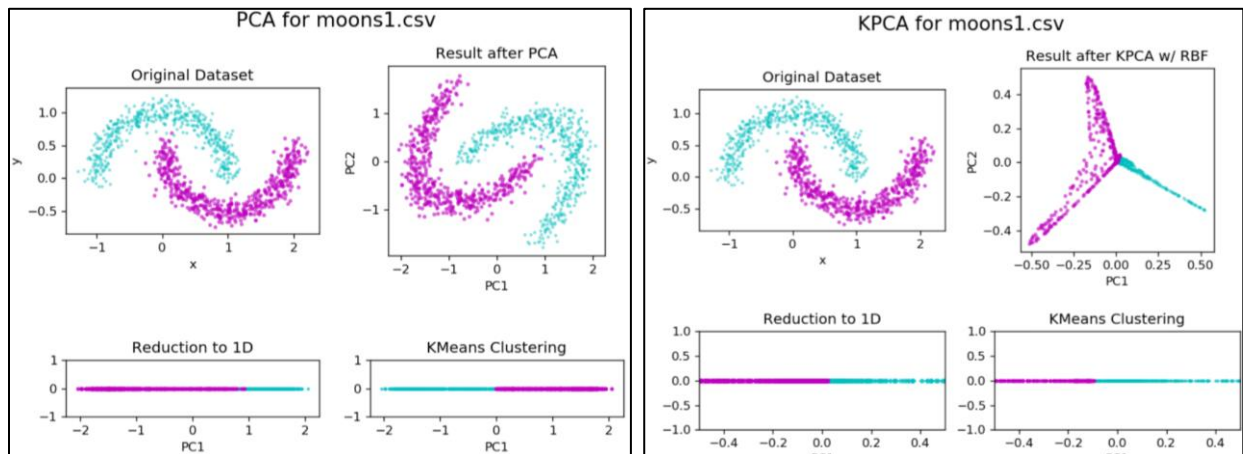
For the moons1 dataset, a gamma value of 27 was chosen. After applying dimensionality reduction with PCA, we see for a third time that the accuracy for the PCA Clustering was very poor, clearly seen by the Reduction and Clustering plots (left bottom row) being different and classified the as opposite labels. On the other hand, we can see that the accuracy for the KPCA Clustering was high, as seen by the similarity between the KPCA's Reduction and Clustering plots (right bottom row), with the only difference being the division point.

For the spiral dataset, a gamma value of 38 was chosen. After applying dimensionality reduction with PCA, we see for a fourth time that the accuracy for the PCA Clustering was very poor, clearly seen by the Reduction and Clustering plots (left bottom row) being different and classified the as opposite labels. On the other hand, we can see that the accuracy for the KPCA Clustering was high, as seen by the similarity between the KPCA's Reduction and Clustering plots (bottom row), with the only difference being the division point. The accuracy is not as high as the previous dataset, moons1, since the division points are further apart.