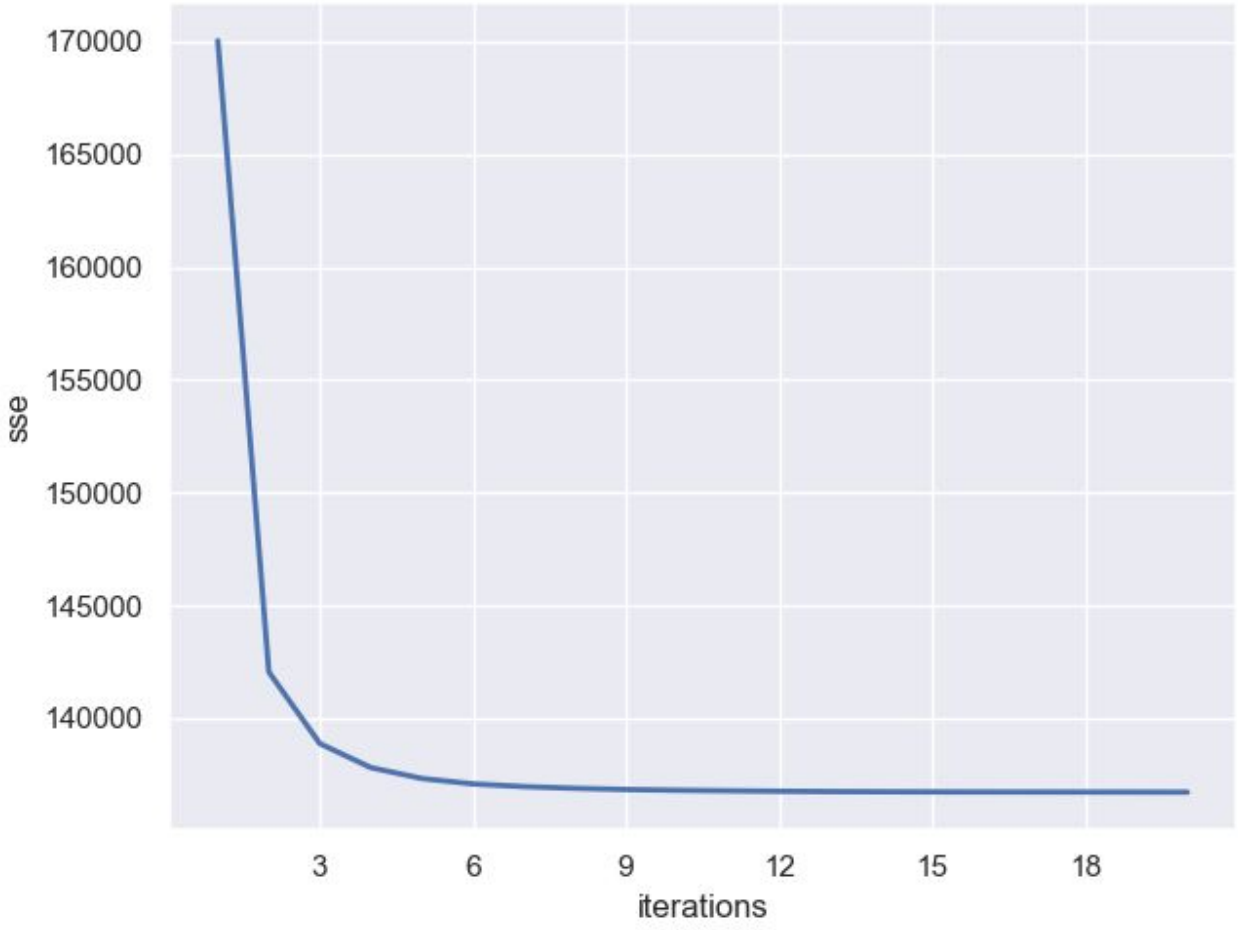
Implementation Assignment 4

Submission for Grayland Lunn, Race Stewart and Joshua Diedrich

Part 1

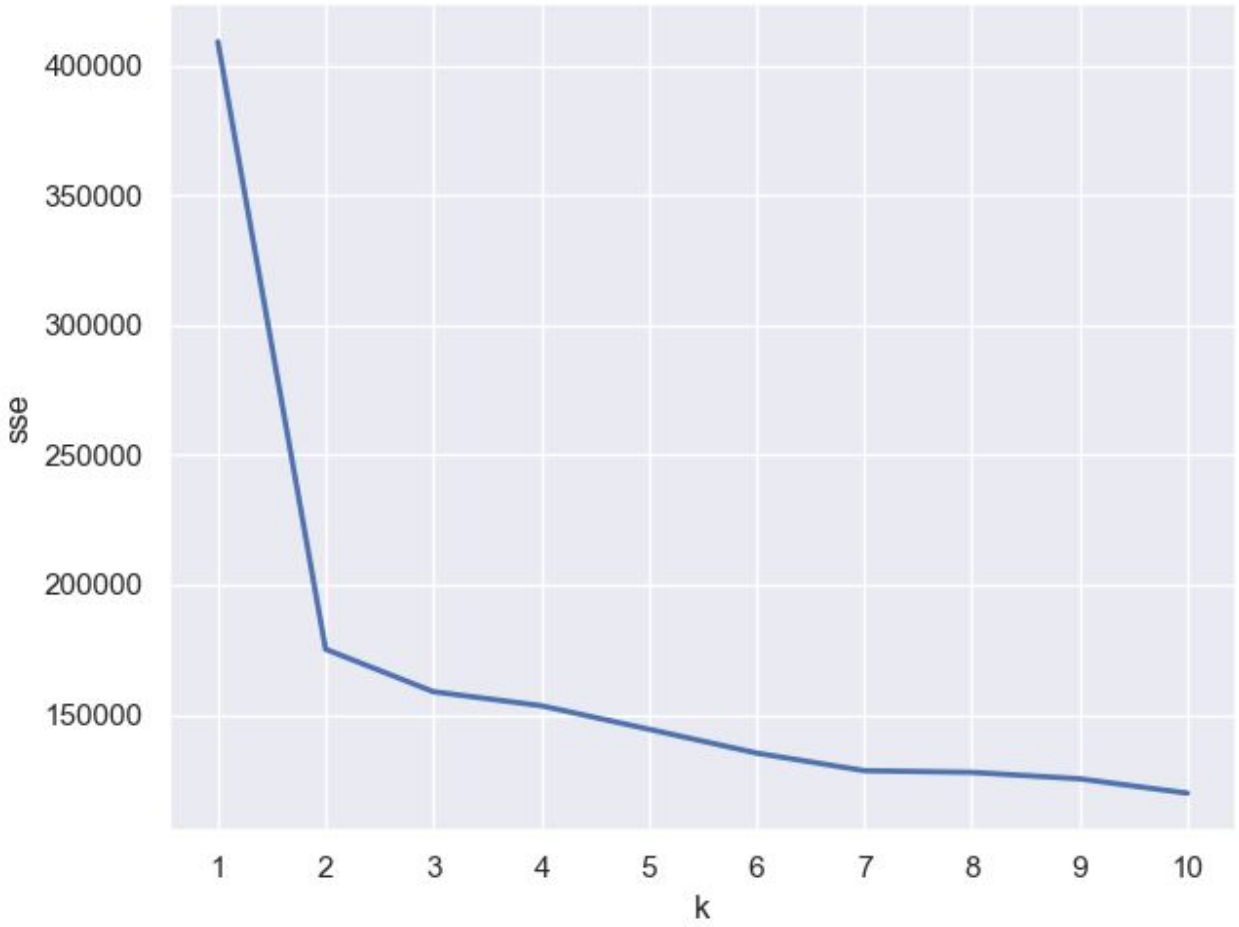
SSE vs Iterations



Plot of sum of square error average over 5 runs vs iterations

Here we see a steady decrease in the sse as the number of iterations increases, in all of the graphs that we saw of sse and iterations, there was never an increase in sse with more iterations. This is a sign that the algorithm is performing correctly. With the average seen here, it looks like stopping after 6-9 iterations would be good enough to reach a final form of the classifier for k = 6, however some runs take longer to decrease because of potential local minima in the sse calculation.

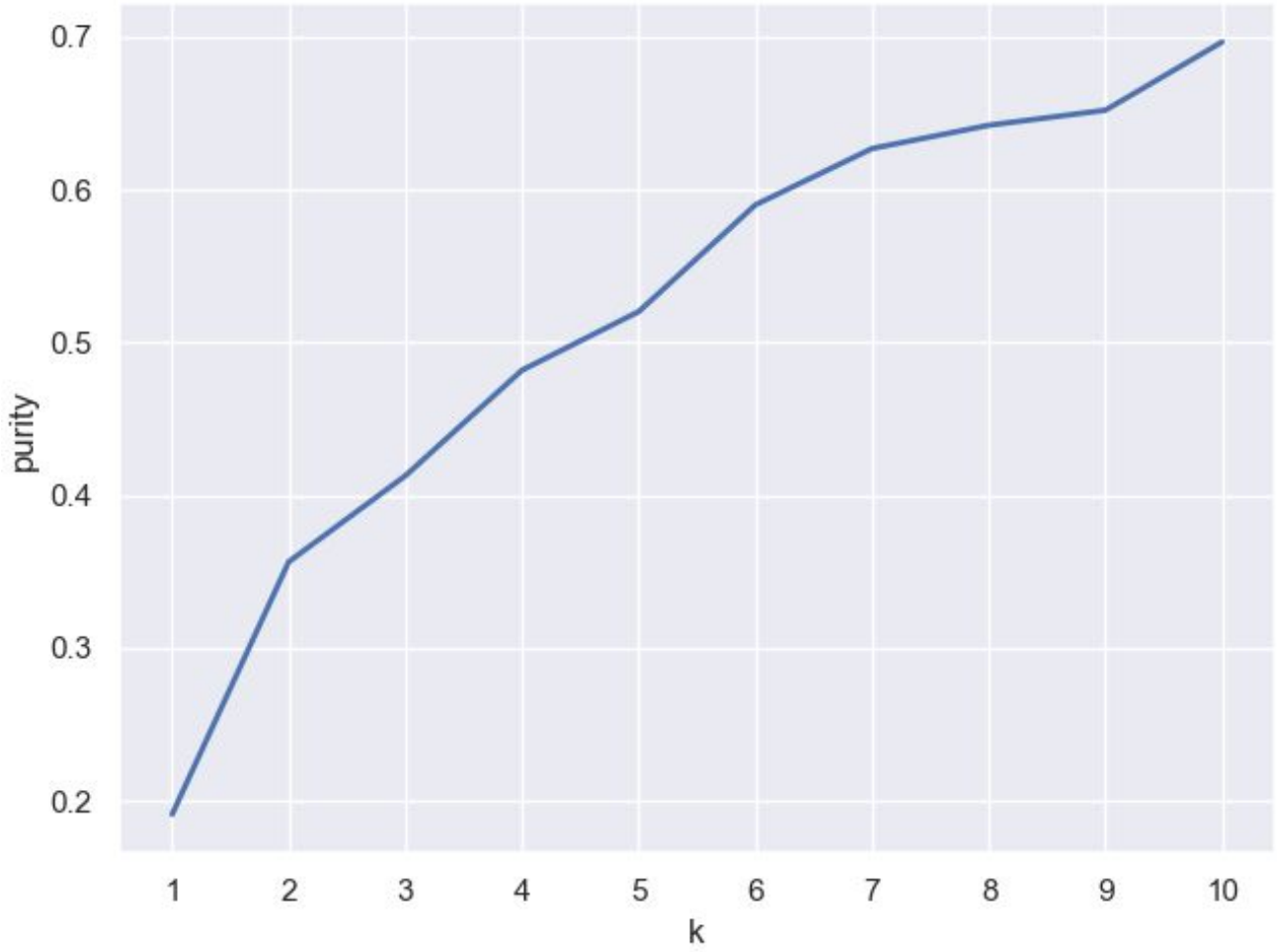
SSE vs K



Plot of sum of square error average over 5 runs for k = 1…10

Here we can see a clear “elbow” at k = 2 classes. While there are 6 classes in real life, it seems that the data only really supports 2. Note that there is very little improvement (percentage wise) in the sse after 2.

Purity vs K

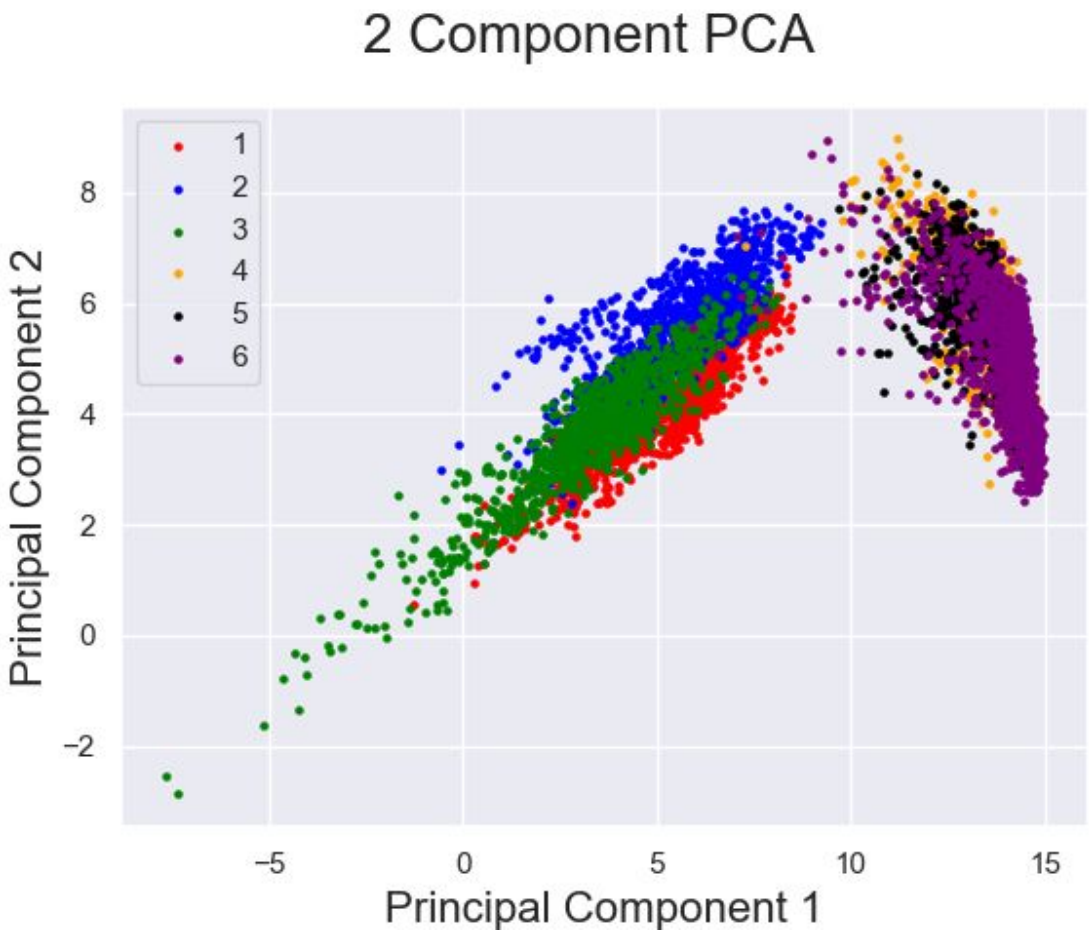


Plot of purity average over 5 runs for k = 1…10

Here we see a steady increase in the purity as k increases. This is expected with this algorithm, as the worst case for an additional class would be classifying no extra points, and the worst case is the same as the previous case. The elbow/knee of this graph can also be seen at k=2, however the purity gains are much more pronounced with a larger k.

Part 2

Visualized Data



Above is the plot of the six classes after PCA is applied. This graph visualizes the first two eigenvectors being transformed onto the data. The default retain ratio of 0.9 was used.

Best Number of Dimensions

With a retain ratio of r, we saw the value of d (where d is the number of dimensions) get to 33. This means after 33 dimensions, we are just under 90% of the original variance from 561 dimensions.

Dimension Reduction Discussion

Initially, there is little difference between the resulting purities with and without PCA.

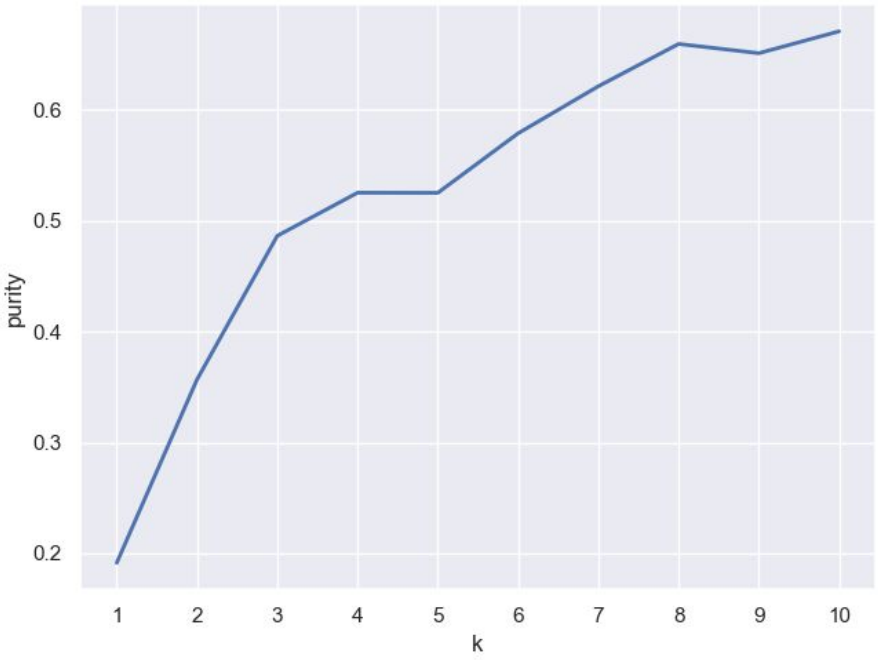


Figure: Purity with no PCA

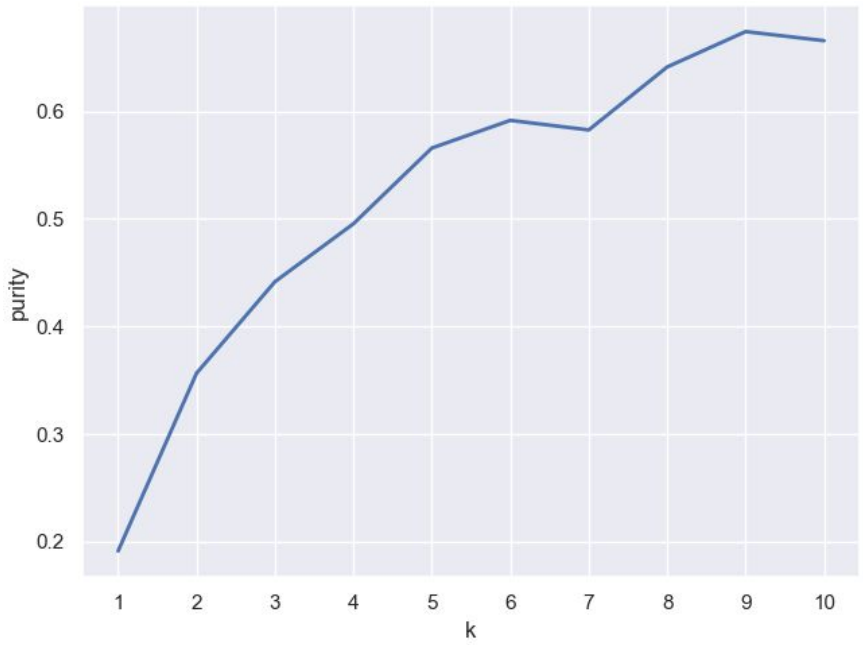


Figure: Purity with PCA and retain ratio 0.9

As you can see above, the purity is slightly decreased by a retain ratio of 0.9. Because of this, we adjusted the retain ratio until we reached a value of 0.95, which gave us slightly better purity while still reducing dimensions.

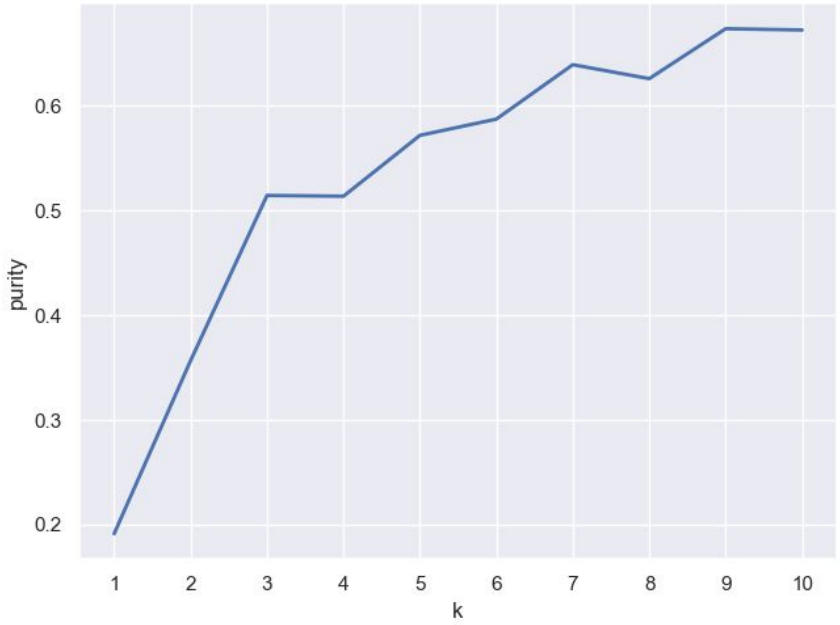


Figure: Purity with PCA and retain ratio 0.95

The retain ratio of 0.95 gave us better purity than 0.9, while remaining close to the purity with no reduction. This leads us to believe 0.95 is the best retain ratio for our data.