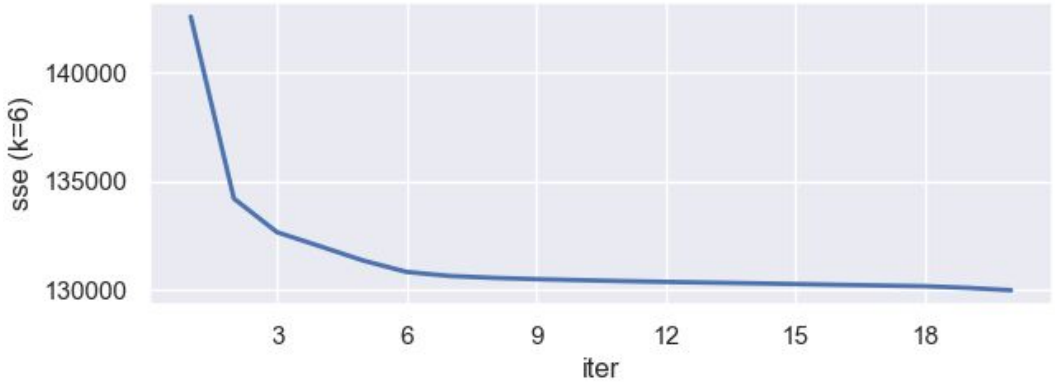
Evan Hopper-Moore, Matthew Jordan
Dr. Xiaoli Fern
CS 434
June 1st, 2020

Assignment 4

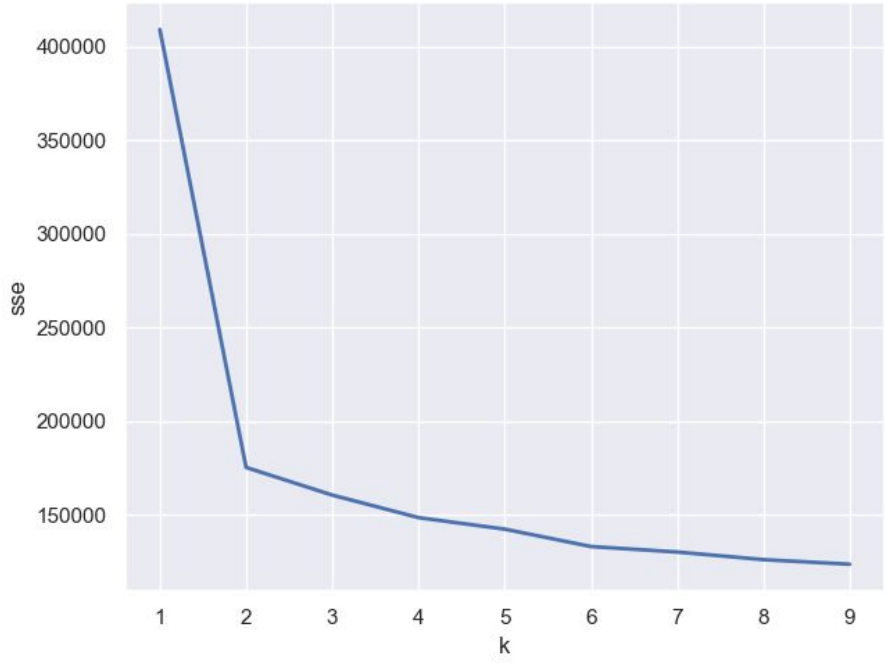
Instruction

Our project doesn't differ from the starter code as far as how to run it. Using python3 and inside of the /src directory, run "python main.py [--pca 0|1] [--kmeans 0|1]" to run the project with pca or kmeans. The default options runs the k-means algorithm with default settings but the parameters can be set with options pca_retain_ratio, kmeans_max_k, kmeans_max_iter, and root_dir.

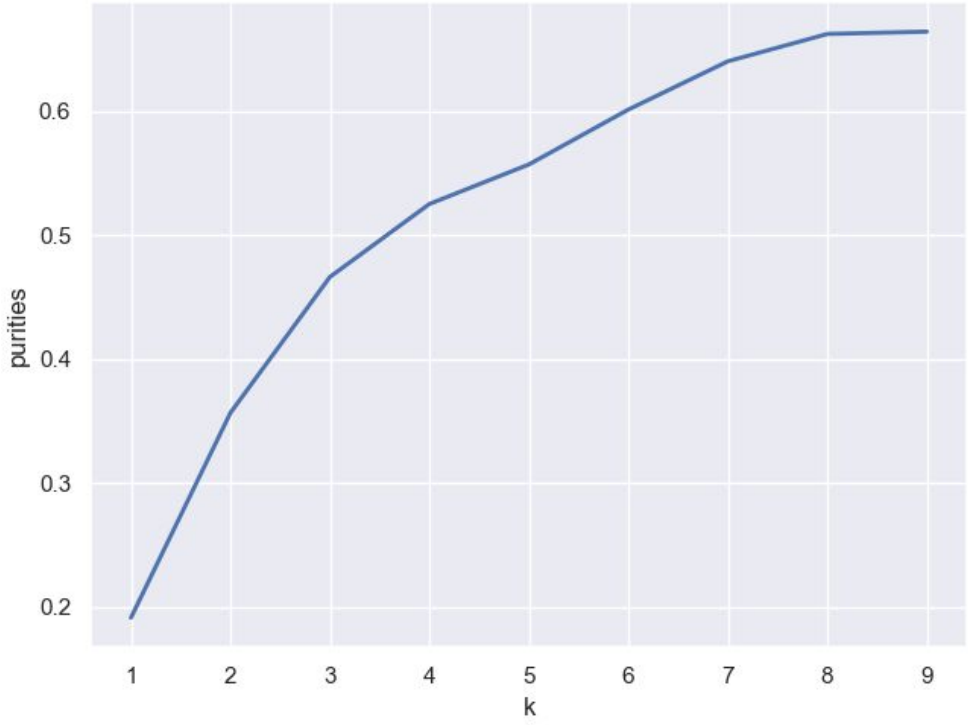
1. The graphs and analysis of the K-Means algorithm are shown below.
    a. The average SSE (averaged over 5 runs of k-means) for k=6 over the iterations is plotted below. As the algorithm runs more iterations, the SSE reduces quickly at first, then flattens out near what must be the minimum SSE possible.



    b. The average SSE for k = 1, 2, ...10 is shown below. The elbow of the curve is at k=4 which means this must be the best k.

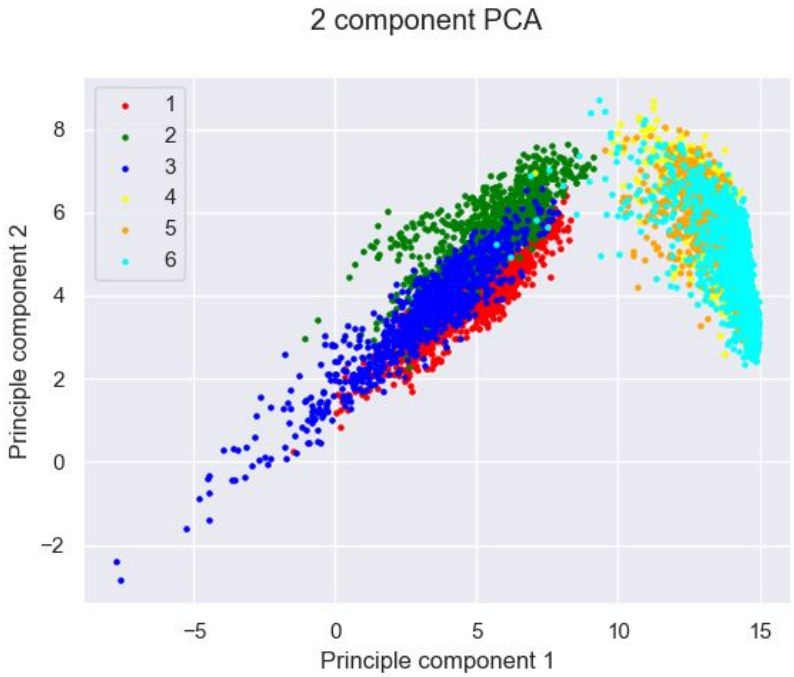


c. The average purity (over 5 runs of k-means) for k = 1, 2, ...10 is shown below. The highest purities are reached at larger values of k, but it also follows the shape of the elbow curve.



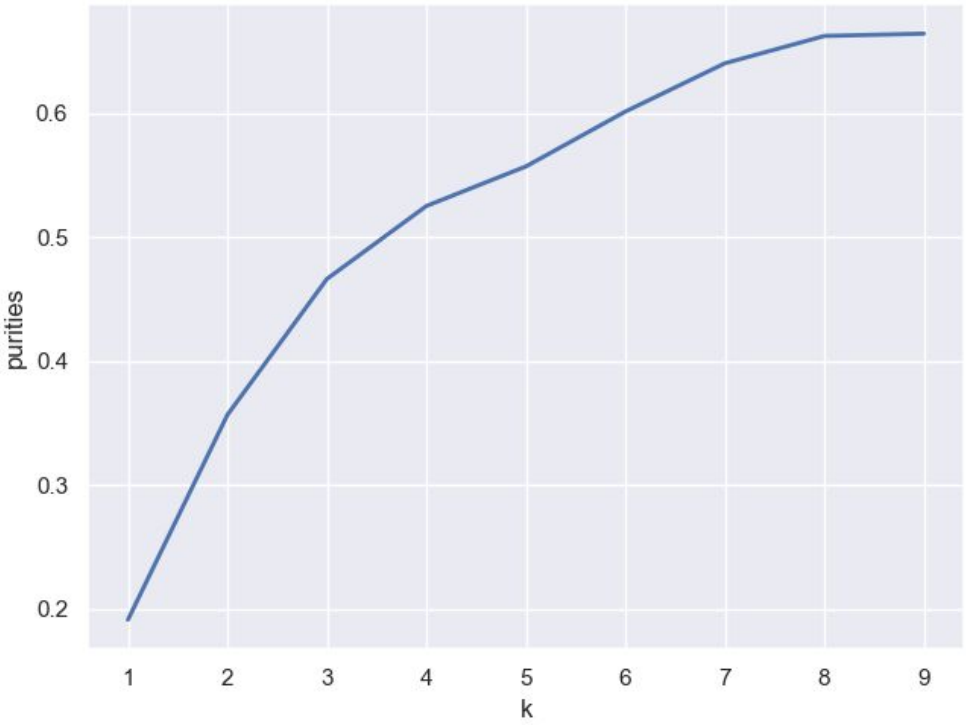
2.

a.

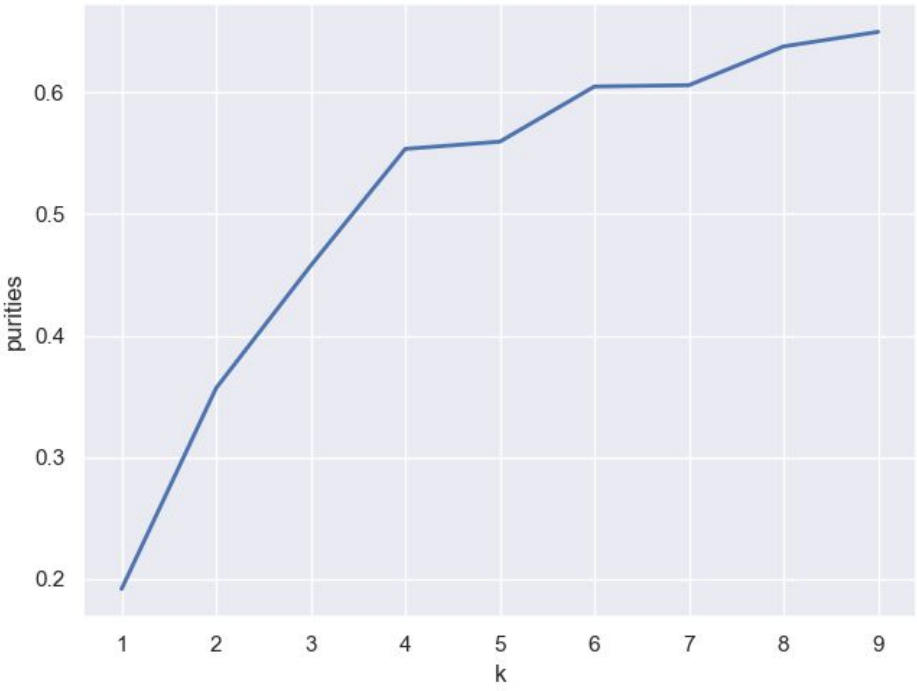


b. This is the visualized data for the 6 classes that were discovered. The ratio of 0.9 was used.

c. This is the purity vs k without PCA

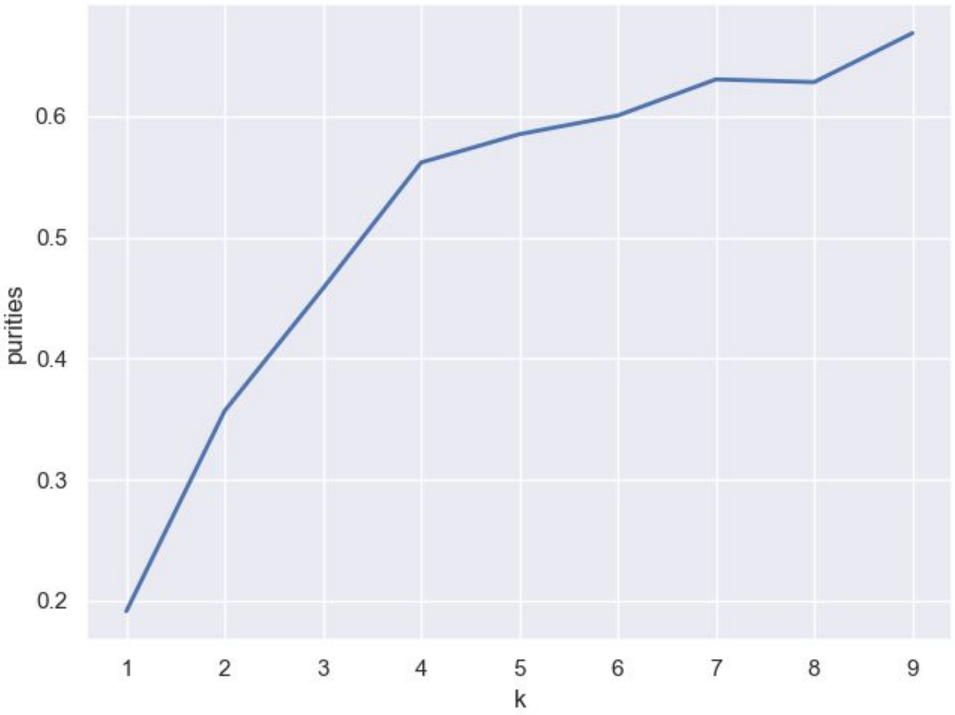


d. This is the purity vs K with PCA with a ratio of 0.9



e. Given the lack of purity in the dimension reduced graph, the ratio was tested to get the best fit between 0.9 and 1 to get a value of 0.95. Using this ratio, we were

able to get the graph below.



f. Given the balance between purity and the reduction, we believe that this is probably the best ratio to use for PCA.