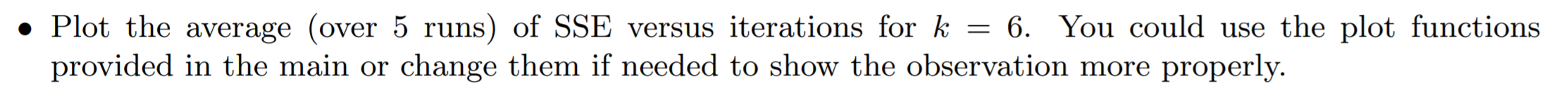
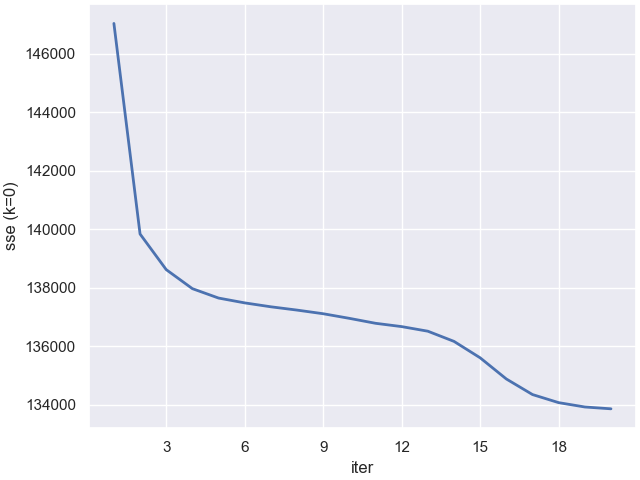
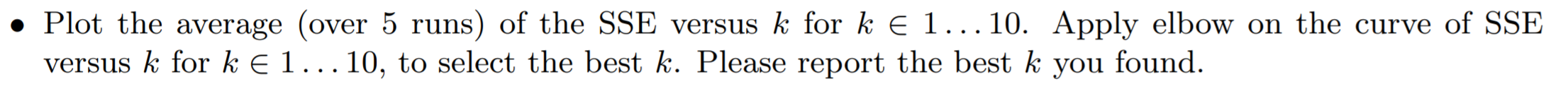
Assignment 4 Report

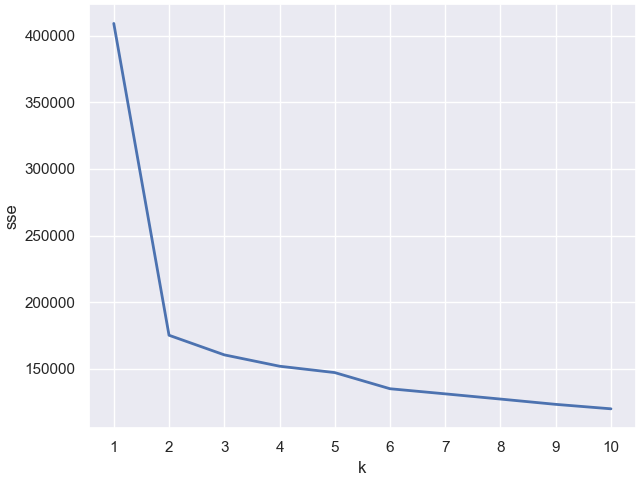




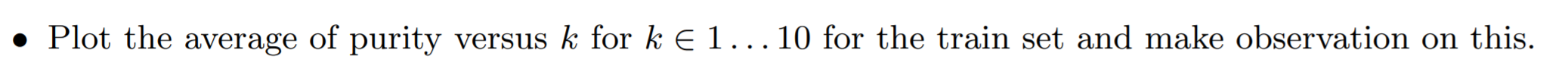
When K=6, the more iteration you do, the less SSE you get. It makes since that SSE will decrease when you have more iterations, which bring you close to the converge.

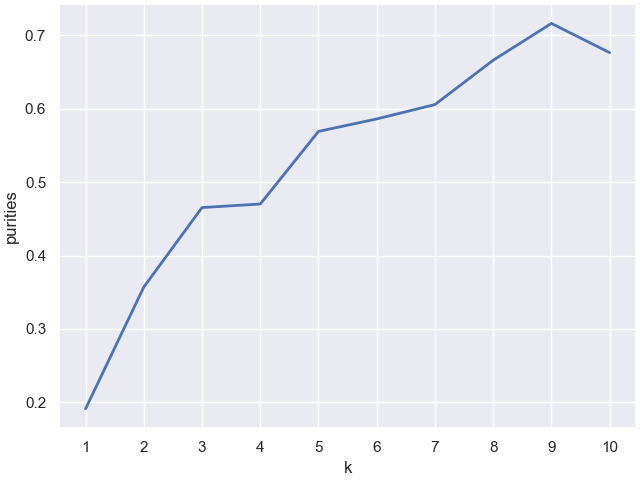
I expect that once iter>20, the curve will be flat since there is no much to be improved.

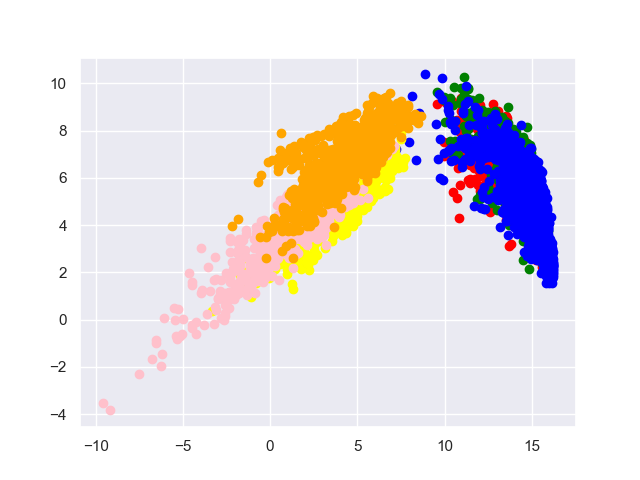
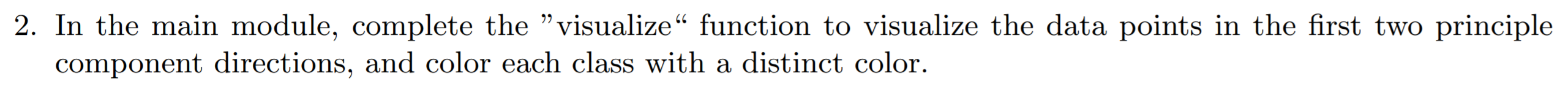




The more clusters you have, the less SSE you get. We can see a clear elbow at K=2, however, this is not good enough for the dataset. Another elbow is at k=6, this one fit our goal well, so I will pick k=6 as the best K I found.

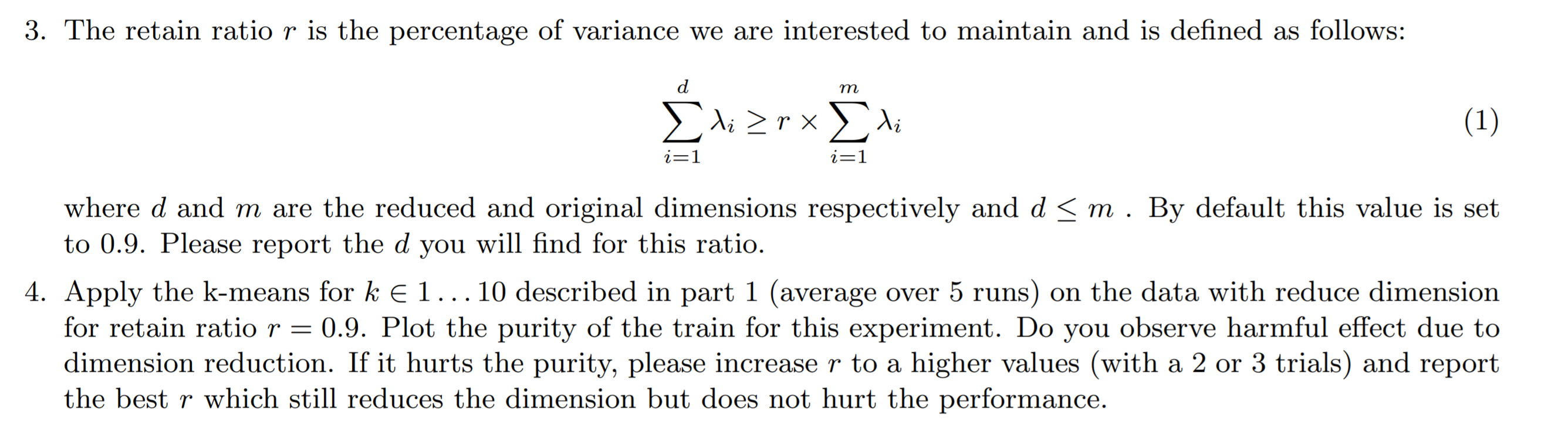


  
Purities increases when K increases until 9, after that Purities decreases. It makes sense that if we have more cluster, there is less data in each cluster, and our purities will increase. However, if we have too many clusters, the purities will be decrease due to too many misclassification



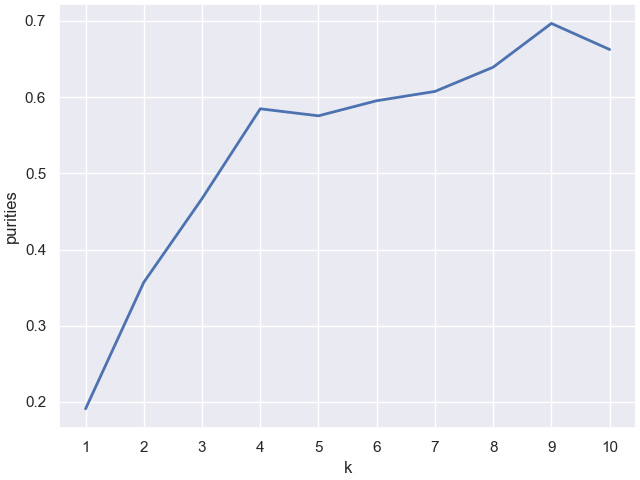
color=['red','green','blue','yellow','pink','orange']

I have 6 colors for 6 different classes. It does not show 6 clear clusters on this 2-D graph, However, we can see that it spread out data well in X-axis, which means that we picked the right principal component direction. We also reduce 541 features to 2 features, so it makes sense that the graph does not show clear cluster.

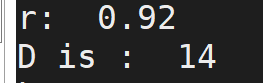




The ‘d’ I found is 34



The best purity is about 0.7 when k=9. Compare to k-means cluster, it does not hurt the purity after dimension reduction.



When r=0.92

I found the max purity is 0.71