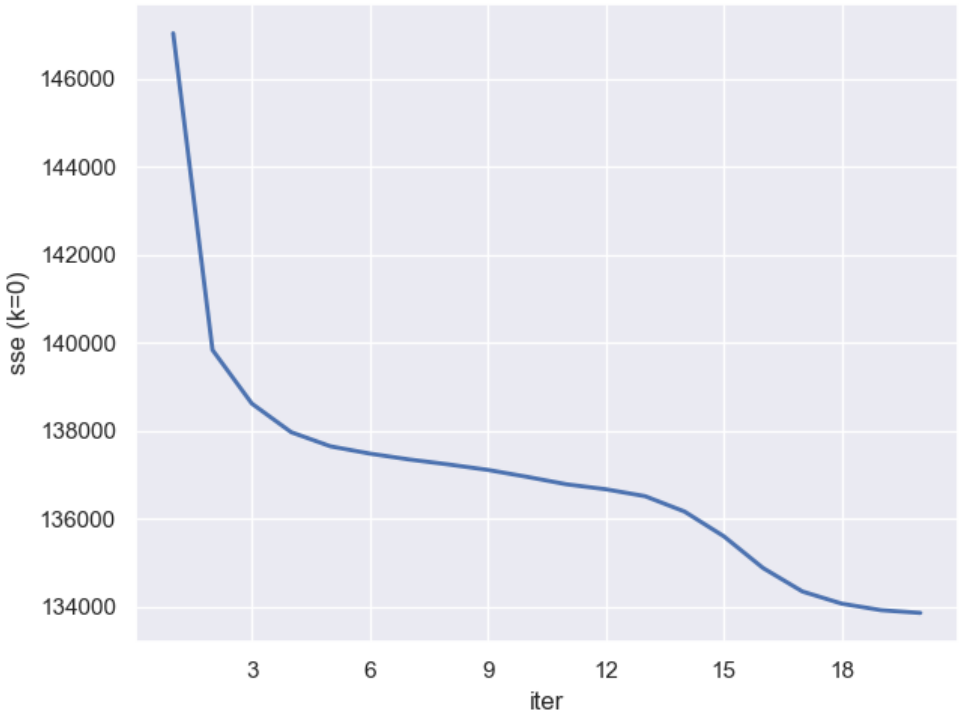
Assignment 4 Report

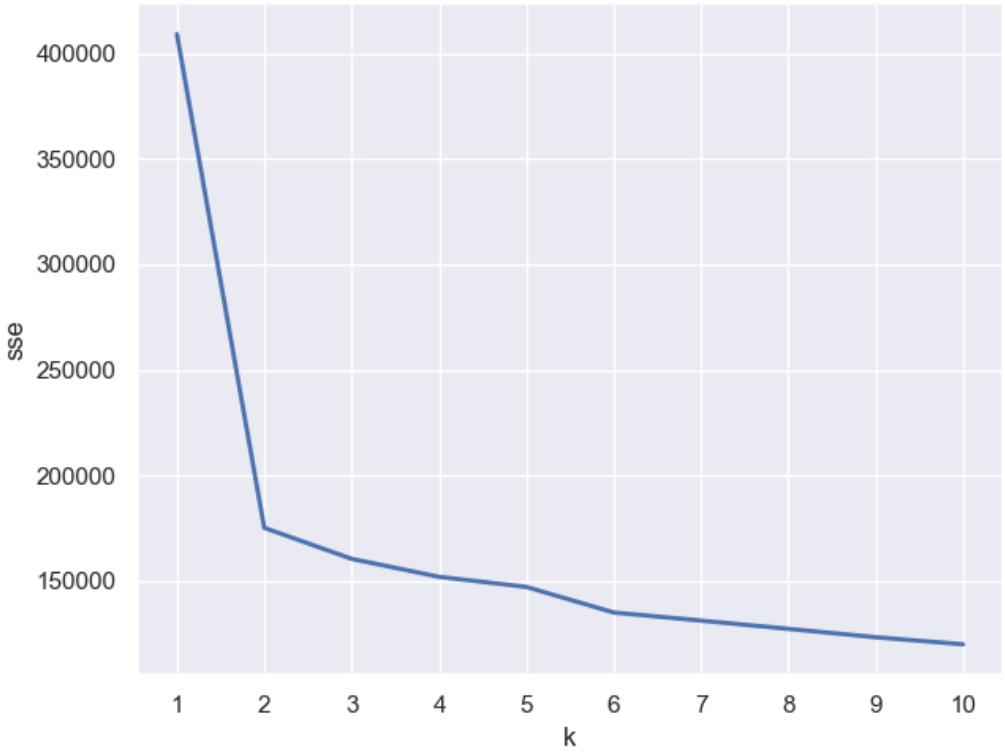
- Plot the average (over 5 runs) of SSE versus iterations for $k = 6$. You could use the plot functions provided in the main or change them if needed to show the observation more properly.



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.

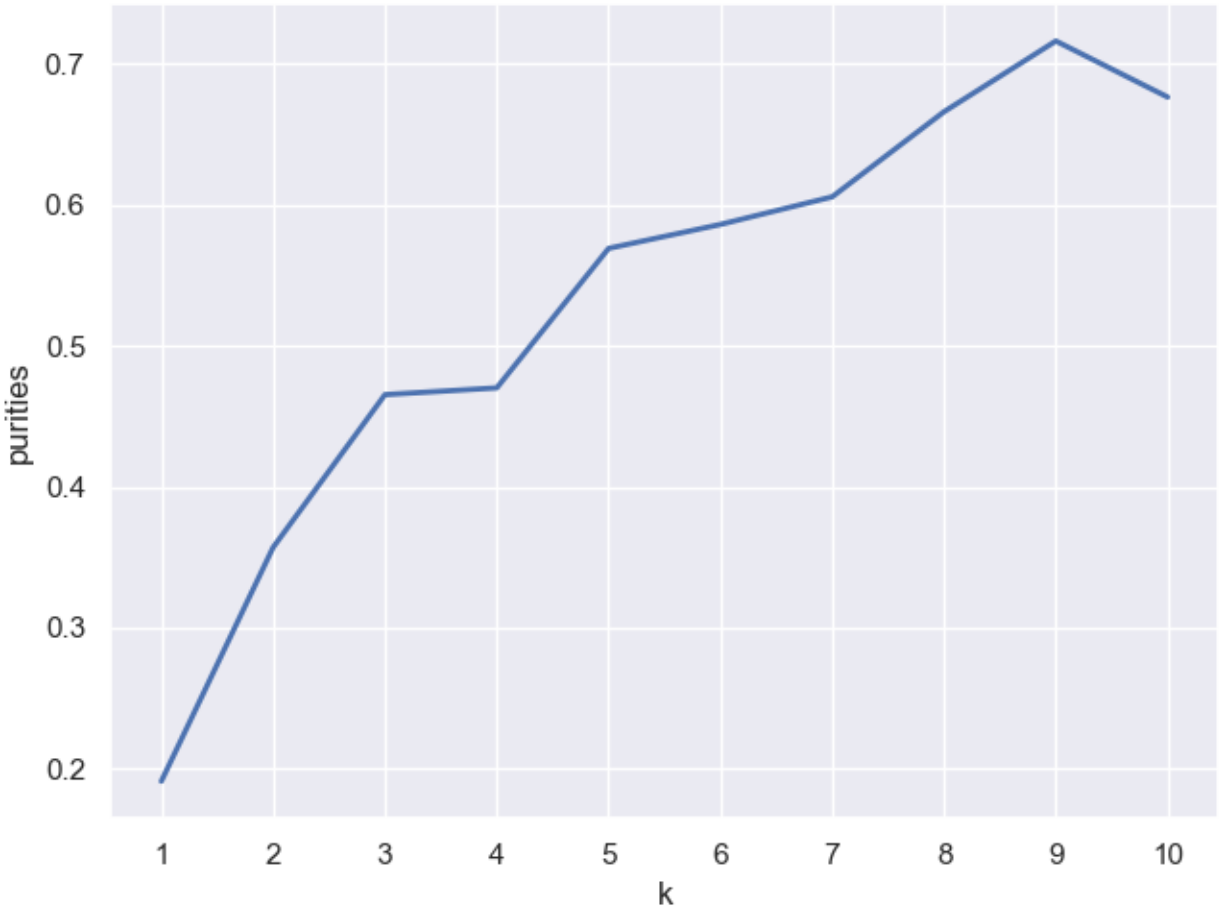
I expect that once iter>20, the curve will be flat since there is no much to be improved.

- Plot the average (over 5 runs) of the SSE versus k for $k \in 1 \ldots 10$. Apply elbow on the curve of SSE versus k for $k \in 1 \ldots 10$, to select the best k. Please report the best k you found.



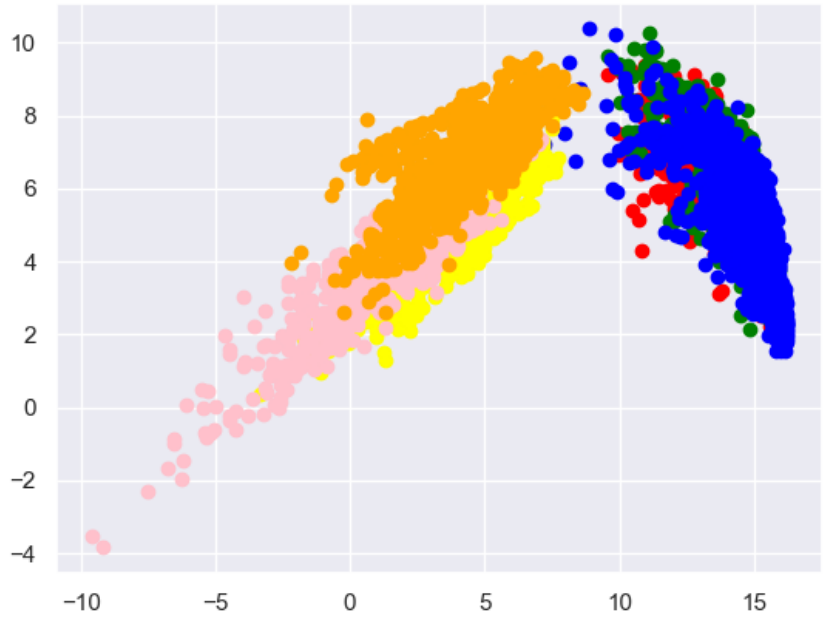
The more clusters you have, the less SSE you get. We can see a clear elbow at K=2, however, this is not good enough for the dataset. Another elbow is at k=6, this one fit our goal well, so I will pick k=6 as the best K I found.

- Plot the average of purity versus k for $k \in 1 \ldots 10$ for the train set and make observation on this.



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

2. In the main module, complete the "visualize" function to visualize the data points in the first two principle component directions, and color each class with a distinct color.



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.

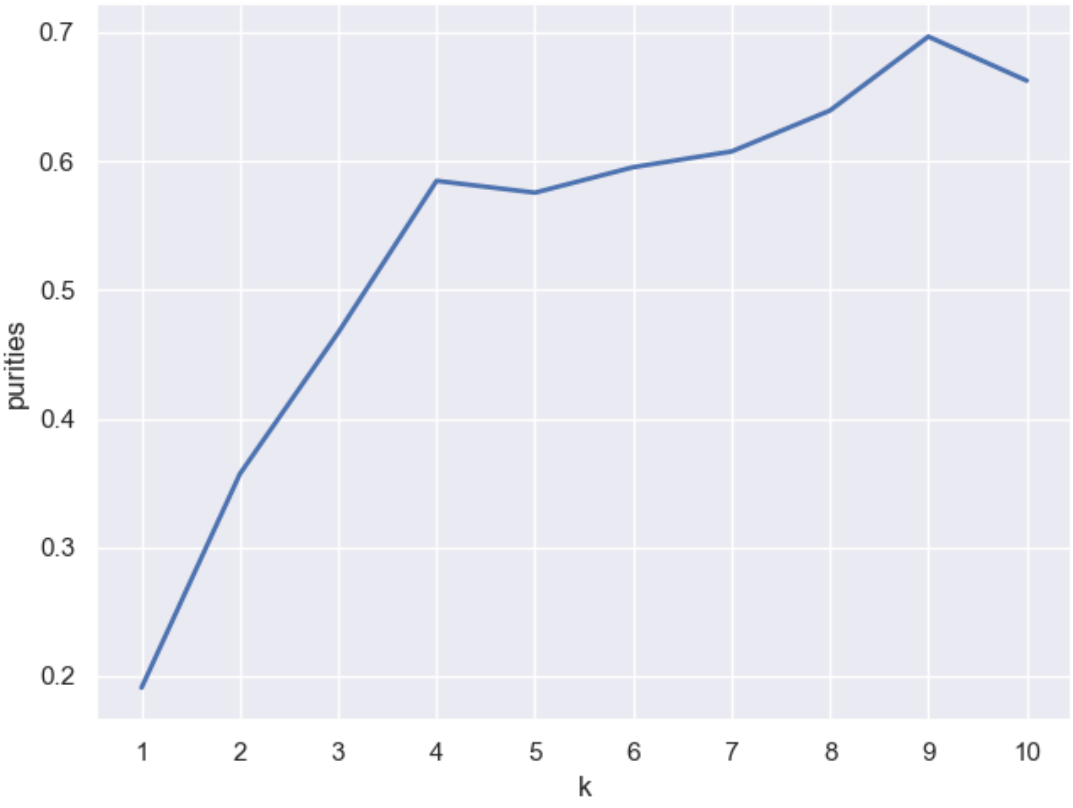
3. The retain ratio r is the percentage of variance we are interested to maintain and is defined as follows:

$$\sum_{i=1}^{d} \lambda_i \geq r \times \sum_{i=1}^{m} \lambda_i \tag{1}$$

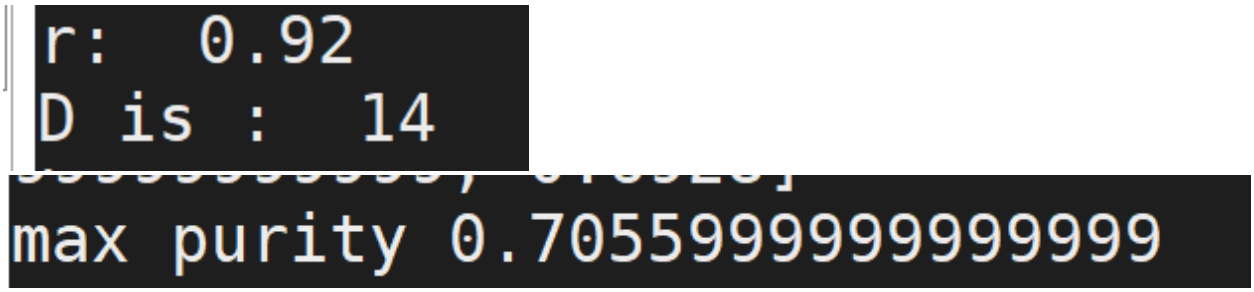
where d and m are the reduced and original dimensions respectively and $d \leq m$. By default this value is set to 0.9. Please report the d you will find for this ratio.

4. Apply the k-means for $k \in 1 \ldots 10$ described in part 1 (average over 5 runs) on the data with reduce dimension for retain ratio $r = 0.9$. Plot the purity of the train for this experiment. Do you observe harmful effect due to dimension reduction. If it hurts the purity, please increase r to a higher values (with a 2 or 3 trials) and report the best r which still reduces the dimension but does not hurt the performance.

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