Samar Ibrahim Antar

Assignment#7 :Clustering and Dimensionality Reduction

- **PCA() function:**

- It takes dataset and number of components.
- First calculate mean to center the data.
- Calculate the covariance matrix of the mean-centered data.
- Calculate Eigenvalues and Eigenvectors of the covariance matrix.
- np.argsort returns an array of indices of the same shape, then sort the eigenvalues in descending order.
- similarly sort the eigenvectors.
- Select a subset from the rearranged Eigenvalue matrix[num_components].
- Finally, transform the data.

- **K_means_cluster() function:**

- It takes dataset and number of clusters.
- Compute the initial centroids randomly by creating an empty centroid array.
- creating (k=number of clusters) random centroids.
- Then, finding the distance between the points. Euclidean distance is most commonly used for finding the similarity by:
  - creating an empty array > (euclid).
  - finding distance between for each centroid>
    dist=(X-centroids[:,k])**2.
  - then concatenate `euclid` with dist by >>`euclid=np.c_[euclid,dist].`

  -storing the minimum value we have computed in a variable called >> minimum.

- regroup the dataset based on the minimum values and calculate the centroid value by:

  - computing the mean of separated clusters.

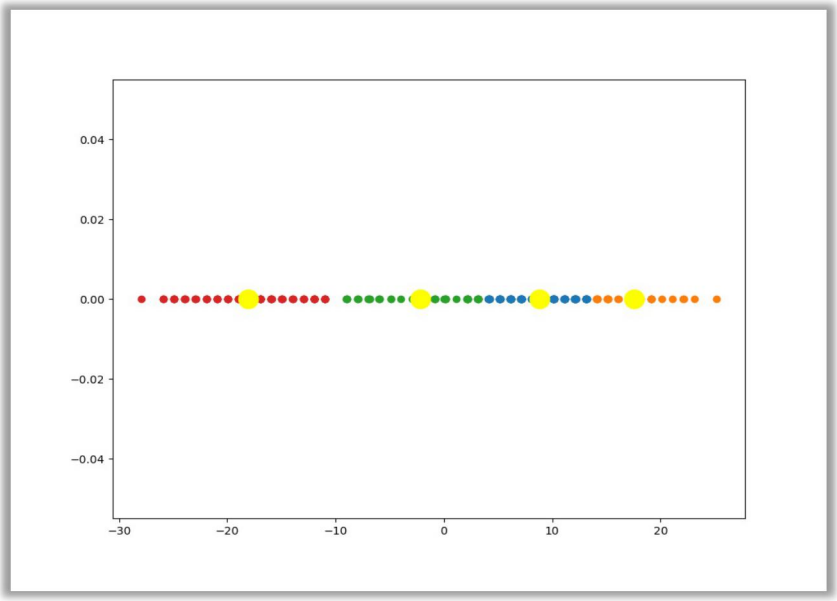
  - assigning of clusters to points.

- computing mean and updating it.

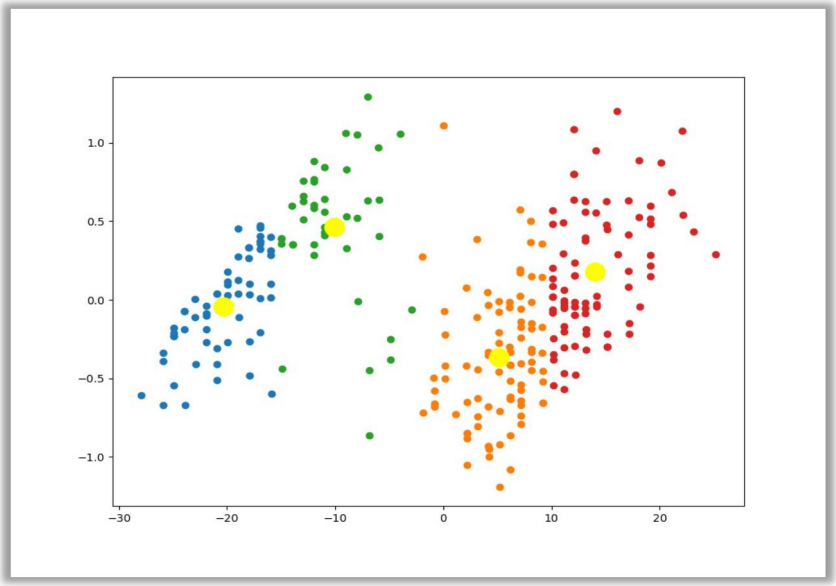
- repeating the above steps again and again until reaching the convergence.

- **Plot() function**:

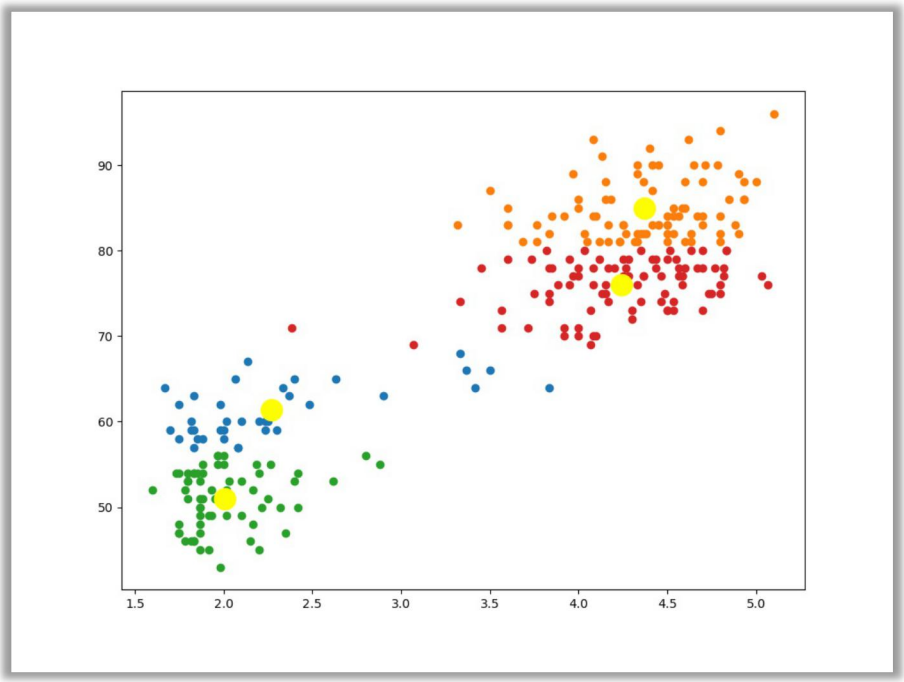
- to plot one component of PCA after entering to k_means_cluster()function.



- plot two components of PCA after entering to k_means_cluster()function.



- And Plot data before PCA:



- So, from one_component and two_components of PCA plots: Concludes that The increase the number of principal component from one to two or more components, this leads to minimum variance in features and thus almost increasing the misclassification (decreasing the separations between the classes) in data it represents until reaching the representation of dataset like in (plot data before PCA).