Stat 6559 / 4560: Statistical Exploration of Data
Assignment #3: Due Monday April 10, 2023 in class

Question 1 a) Generate a simulated data set with 50 observations in each of three classes (i.e. 150 observations total), and 50 variables.
Be sure to add a mean shift to the observations in each class so that there are three distinct classes and choose covariance matrix of your choice.
(b) Perform PCA on the 150 observations and plot the first two principal component score vectors. Use a different color to indicate the observations in each of the three classes. If the three classes appear separated in this plot, If not, then return to part (a) and modify the simulation so that there is greater separation between the three classes.
c) Use diagnostic tools such as variance contribution plot and offer your comments.
d) Also use loading plots & scatter plots of for first two PCs and interpret the results

Question 2. a) Choose any photo of your choice and convert it into the numerical data. Perform a PCA and reconstruct back the data based on first 100 PCA. Compare the file sizes of original photo and new photo based on PCA. Comment on the quality of photo based on PCA.

b) i) For matrix completion, the codes used in the book "Statistical Learning using R" used the svd() function in R. Instead we can use the prcomp(0). write a function using prcomp() for matrix completion. ii) Use the generated data from Q.1(a). Consider only first 10 variables, so that your new data is of size 150 x 10. Randomly select 5 observations from this new data set and assume that these 5 observations are missing. Use Matrix Completion algorithm to estimate missing values using PCA. Offer your comments on the performance of the missing value estimation using PCA.

Question 3. For this problem, use the data in Q1(a).
a) Perform K-means clustering of the observations with K = 3. How well do the clusters that you obtained in K-means clustering compare to the true class labels?
Hint: You can use the table() function in R to compare the true class labels to the class labels obtained by clustering. Be careful how you interpret the results: K-means clustering will arbitrarily number the clusters, so you cannot simply check whether the true class labels and clustering labels are the same. (b) Perform K-means clustering with K = 2. Describe your results.
(c) Now perform K-means clustering with K = 4, and describe your results.
(d) Now perform K-means clustering with K = 3 on the first two principal component score vectors, rather than on the raw data. That is, perform K-means clustering on the $150x2$ matrix of which the first column is the first principal component score vector, and the second column is the second principal component score vector. Comment on the results.
(g) Using the scale() function, perform K-means clustering with K = 3 on the data after scaling each variable to have standard deviation one. How do these results compare to those obtained in Q1? Explain.