HW 3
Machine Learning
PCA (principal component analysis)

HW 3 Problem Statement (provided by Dr. Han Hu, MEEG 491V/591V-028, Fall 2021)
Problem 3-1 Principal Component Analysis and Clustering of Images:
HW-3 uses the same data set as Problem 1-2 of HW-1.

a. Run single value decomposition (SVD) or principal component analysis (PCA) of the images and plot the percentage explained variance vs. the number of principal components (PC).
b. Pick a representative image, run PCA and plot the reconstructed images using different number of PCs (e.g. using PC1, PCs 1-2, PCs 1-10, PCs 1-20, etc.).
c. Calculate the error of the reconstructed images relative to the original image and plot the error as a function of number of PCs.
d. Run a clustering analysis of the boiling images using the PCs (the number of PCs to use is up to your choice) and evaluate the results of clustering.

Principal Component Analysis
PCA is a dimensionality reduction that is often used to reduce the dimension of the variables of a larger dataset that is compressed to the smaller one which contains most of the information to build an efficient model.[1] PCA represents the data in a new coordinate system in which basis vectors follow modes of greatest variance on the data.

PCA also can use for images. For example, if there is a 256*256 image, then it can be considered as a long 1D vector by concatenating image pixels column by column or row by row which is 65536*1 or 1*65536. The number of 65536 is the dimensionality of the vector space. If there are N images, PCA will be still applicable. The graphical illustration of PCA deduction is shown below in Fig 1.

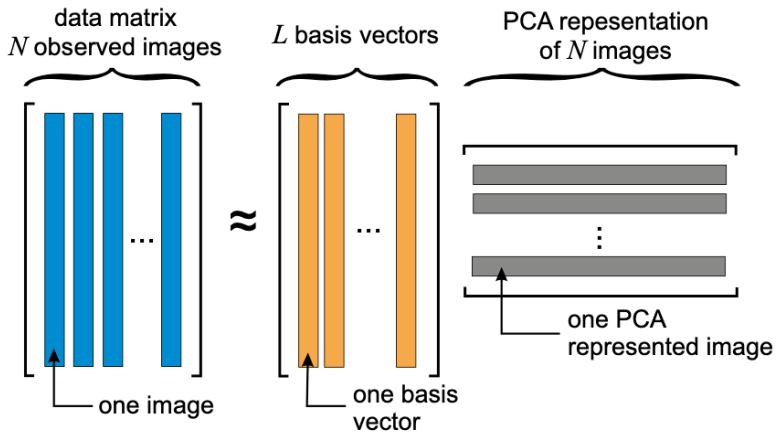

Fig 1. PCA graphical illustration [8]

Code Running Instructions
The environment is default only with scikit-image download.

Coding explanation

The PCA and K-means cluster for images code is attached as a .py file in the zip file. In the first step, I save all the images in the dataset.csv file and label the post-images as 1, the pre-images as 0. I deduct the image size as 240*240 and split the dataset with 80% training dataset and 20% testing dataset.

For the first question, I adapted the code from [3]. For the second question, the code adapted from [9]. The code reference from [10] for question 3. There are several ways to calculate the error but should have the similar shape of the graph. Code of question 4 adapted from. [11]

200 images are used to debug, then all the images are used to train and test the model. It needs to be noticed that results are similar which I attached the graphs in the result section.

Results

Question a). The graphs of percentage explained variance vs. the number of principal components (PC) are shown in Fig 2 and Fig 3 below. 100 principal components are taken.

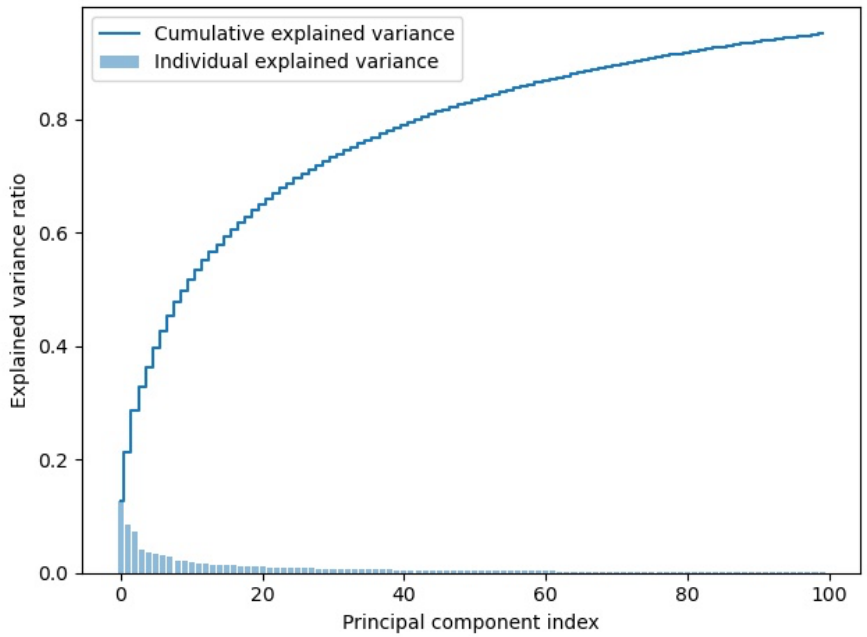

Fig 2. Percentage explained variance vs. the number of principal components (PC): training vs testing dataset: 160 vs 40.

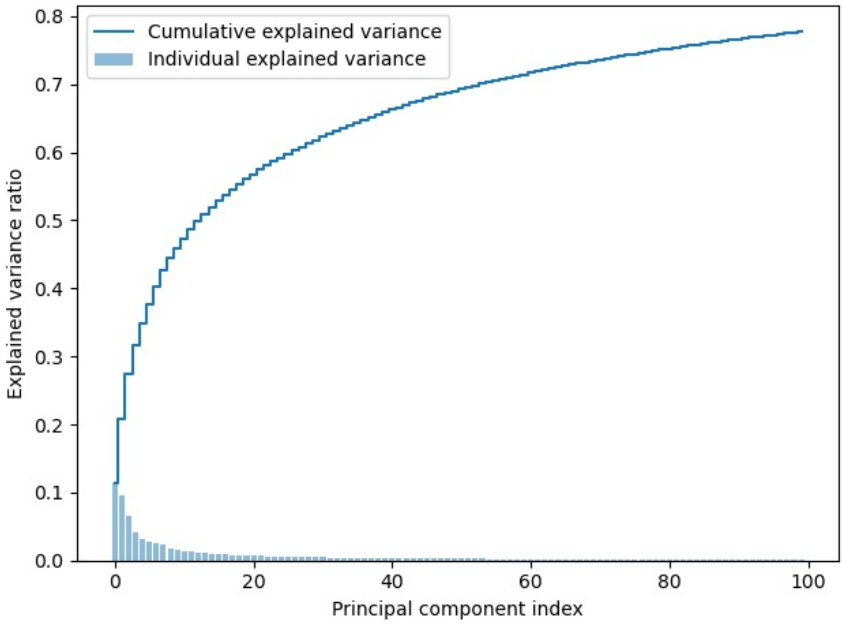

Fig 3. Percentage explained variance vs. the number of principal components (PC): training vs testing dataset: 19112 vs 4778

Question b). The graph of reconstructed images using 1, 2, 10, 20, 50, and 100 PCs by using 200 images are shown below.

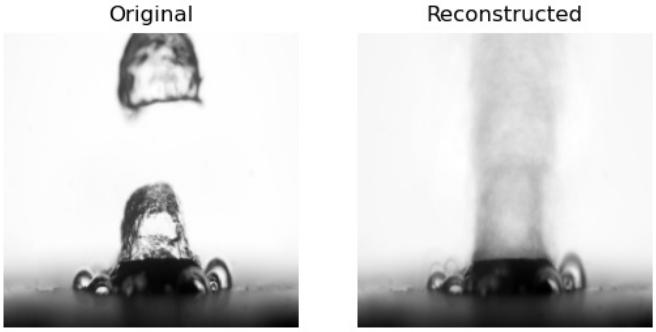

Fig 4. The reconstructed images using PC1.

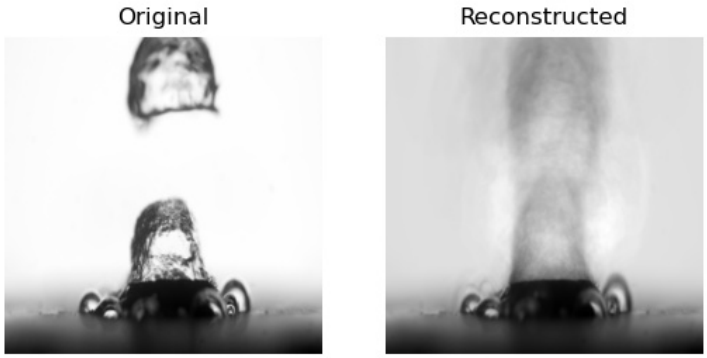

Fig 5. The reconstructed images using PCs 2

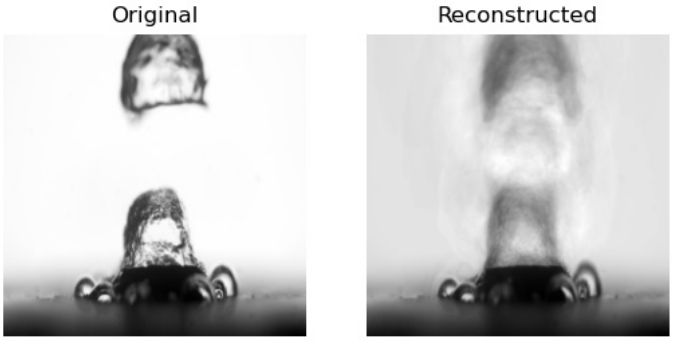

Fig 6. The reconstructed images using PCs 10

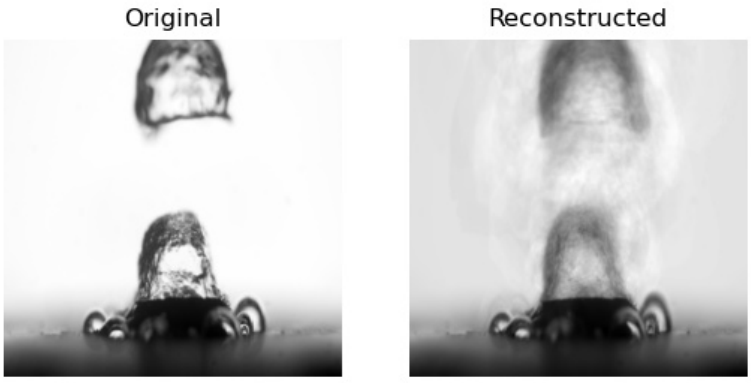

Fig 7. The reconstructed images using PCs 20

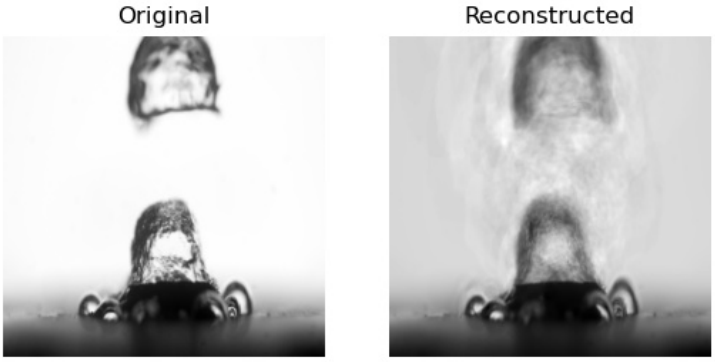

Fig 8. The reconstructed images using PCs 50

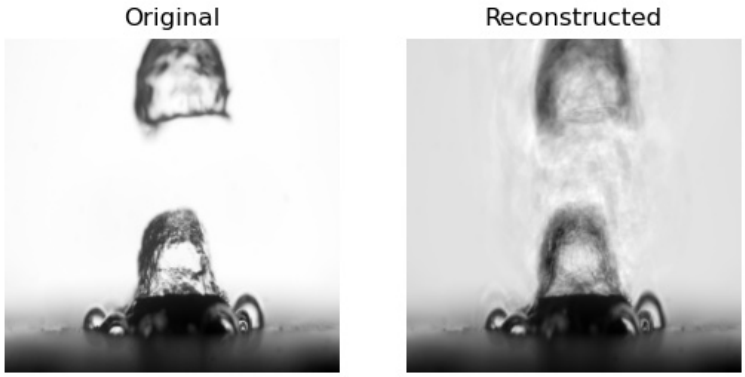

Fig 9. The reconstructed images using PCs 100

The graph of reconstructed images using 1, 2, 10, 20, 50, and 100 PCs by using 23890 images are shown below.

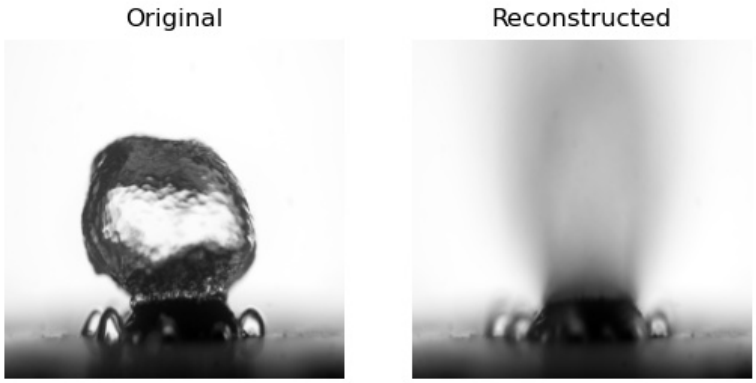

Fig 10. The reconstructed images using PC1

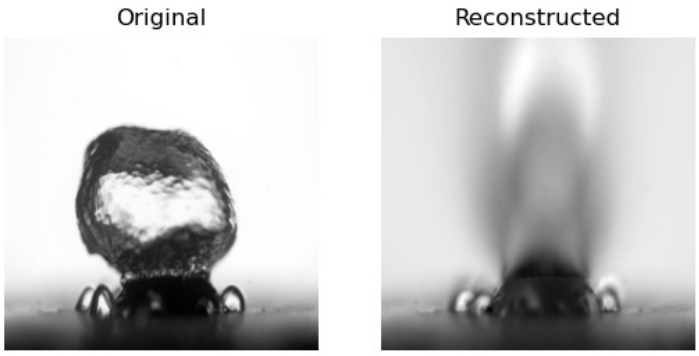

Fig 11. The reconstructed images using PCs 2

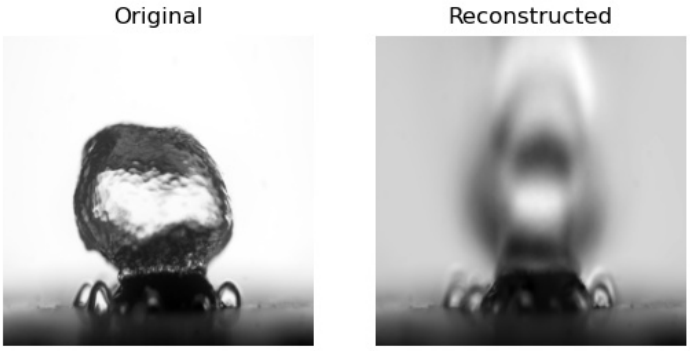

Fig 12. The reconstructed images using PCs 10

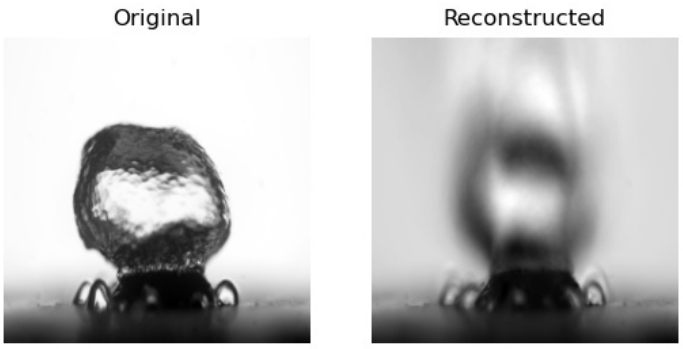

Fig 13. The reconstructed images using PCs 20

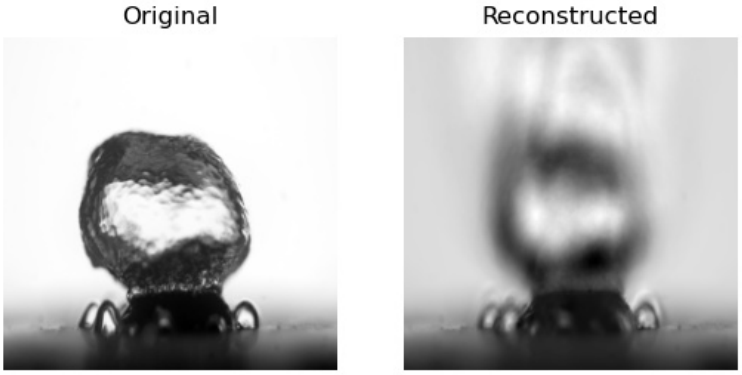

Fig 14. The reconstructed images using PCs 50

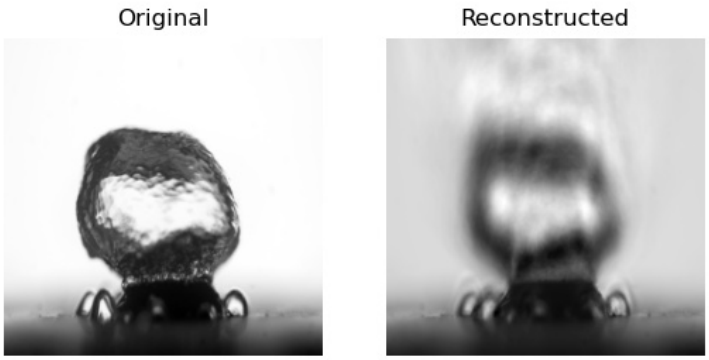

Fig 15. The reconstructed images using PCs 100

The more principal components we used the clearer the reconstructed image will be. That is because the less information lost when the more principal components included.

Question c). The graph of the error as a function of number of PCs 100 by using 200 images are shown below, the x axis is the error, the y axis is the number of PCs.

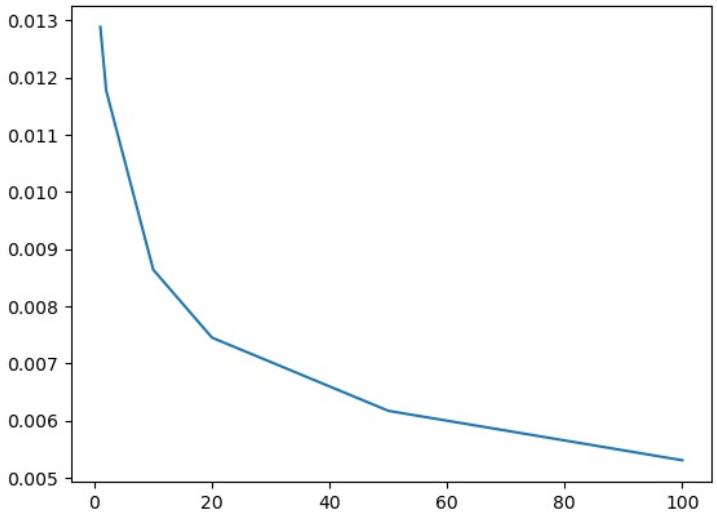

Fig 16. The error as a function of number of PCs 100 by using 200 images

The graph of the error as a function of number of PCs 100 by using 23890 images are shown below, the x axis is the error, the y axis is the number of PCs.

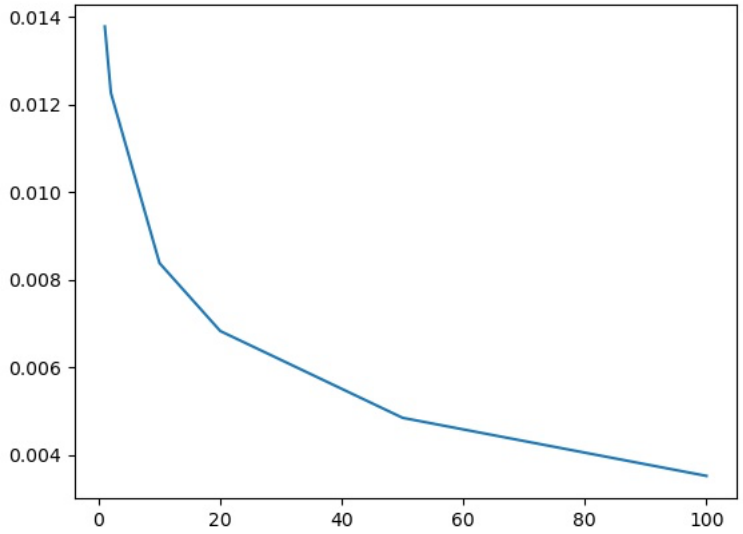

Fig 17. The error as a function of number of PCs 100 by using 23890 images

Question d). K-means clustering analysis of the boiling images using the PCs by 23890 images are shown below.

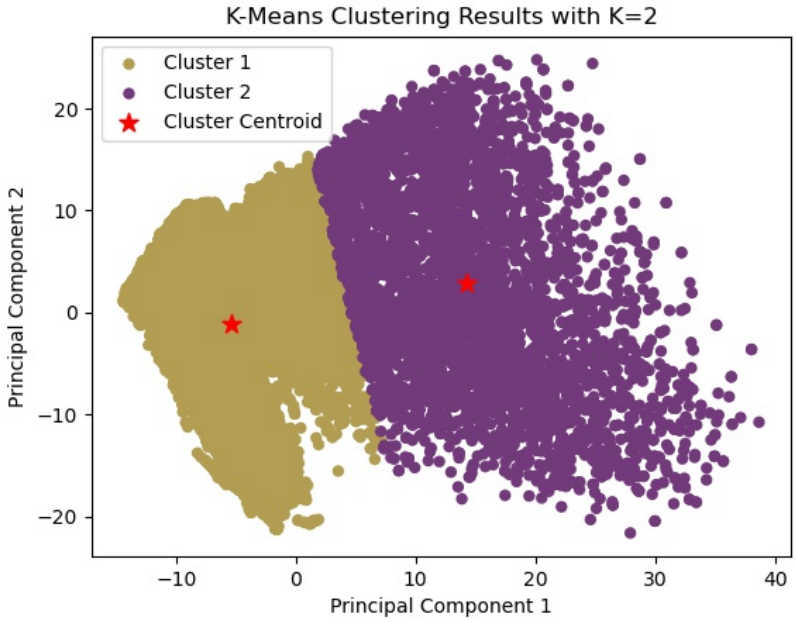

Fig 18. K-means clustering analysis by using 2 PCs

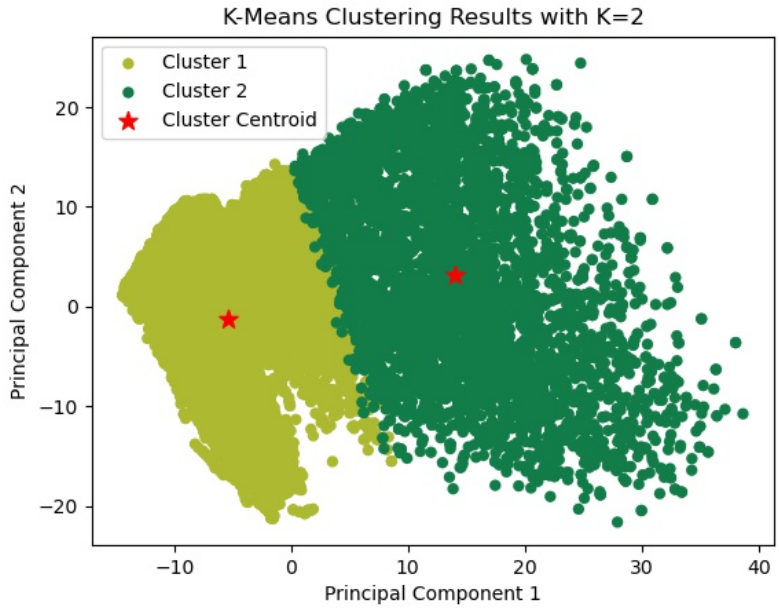

Fig 18. K-means clustering analysis by using 20 PCs

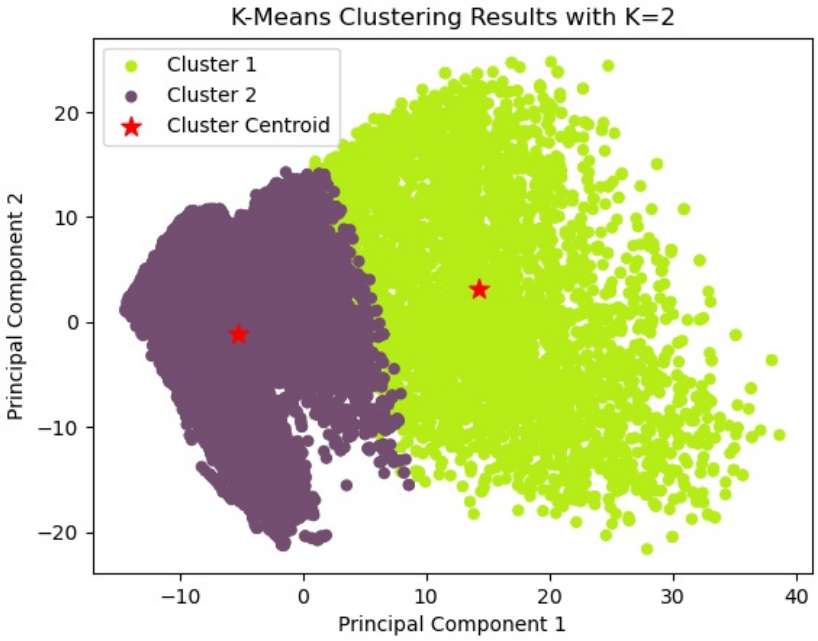

Fig 18. K-means clustering analysis by using 100 PCs

The result of the K-means clustering performs good because it has clear decision boundary when PCs 2, 20, 100 are used.

Challenges

The main challenge for me is to read, save and make a data frame for the datasetÍ.

References

[1] https://analyticsindiamag.com/how-does-pca-dimension-reduction-work-for-images/.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html.

[3] https://vitalflux.com/pca-explained-variance-concept-python-example/.

[4] https://medium.com/@sebastiannorena/pca-principal-components-analysis-applied-to-images-of-faces-d2fc2c083371.

[5] https://scipy-lectures.org/packages/scikit-learn/auto_examples/plot_eigenfaces.html.

[6] https://scikit-image.org/docs/stable/auto_examples/transform/plot_rescale.html.

[7] https://www.youtube.com/watch?v=ZwiDOse1wQU.

[8] http://people.ciirc.cvut.cz/~hlavac/TeachPresEn/11ImageProc/15PCA.pdf.

[9] https://shankarmsy.github.io/posts/pca-sklearn.html

[10] https://www.kaggle.com/ericlikedata/reconstruct-error-of-pca

[11] https://365datascience.com/tutorials/python-tutorials/pca-k-means/