IT350 Assignment 1

NAME: SUYASH CHINTAWAR
ROLL NO.: 191IT109
TOPIC: PCA AND SVD

Note:
1) The colab link has been attached below. After opening the link, if it opens in drive, click on “Open with Google Colaboratory” to view the complete code.
2) Only output screenshots have been attached along with the explanation. Code for the same can be found in the colab notebook.

Colab notebook link:
https://colab.research.google.com/drive/1HRZNn97oRqPFDquIqRLjBwj2XCPyW4I_

Q. Download a dataset of your choice and perform the following.

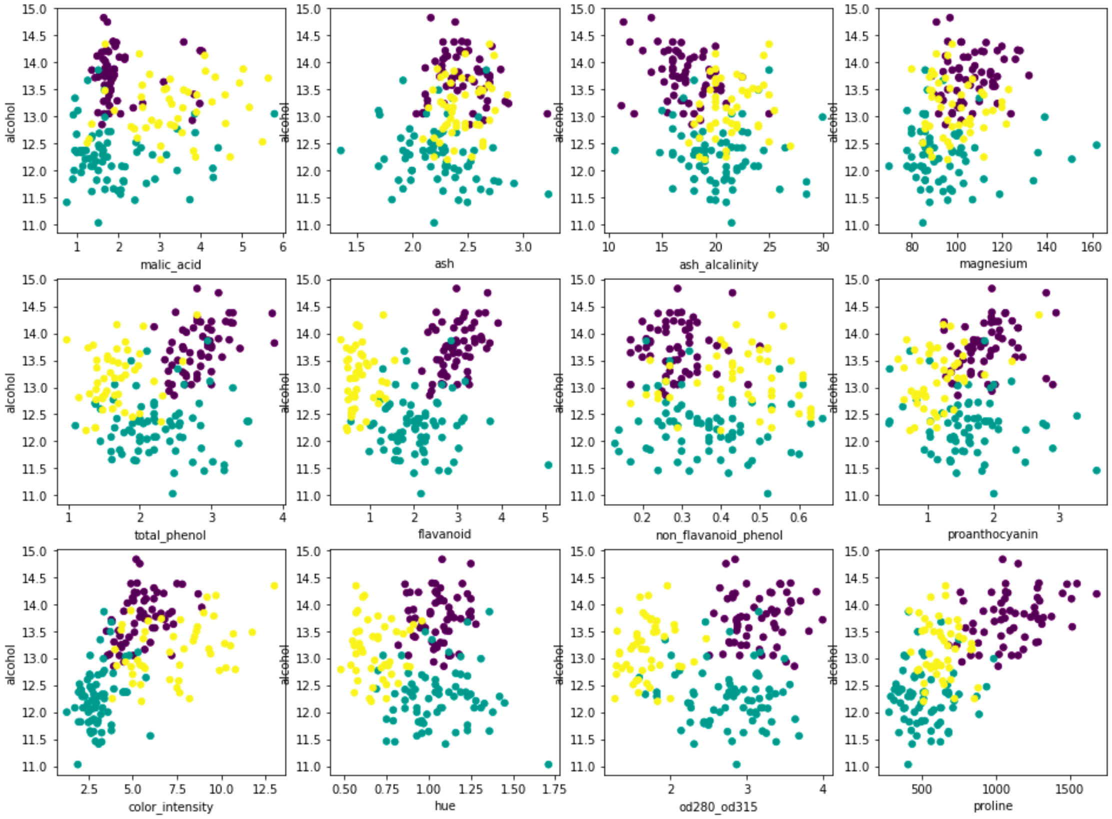
1. Visualize it using multiple dimensions and say why SVD and PCA should be used here.
2. Implement SVD and PCA logic on your own and find the appropriate k-dimensions to represent this data.
3. Visualize the data after applying SVD and PCA.
4. State your conclusions as to how SVD and PCA have helped here.

Solution:

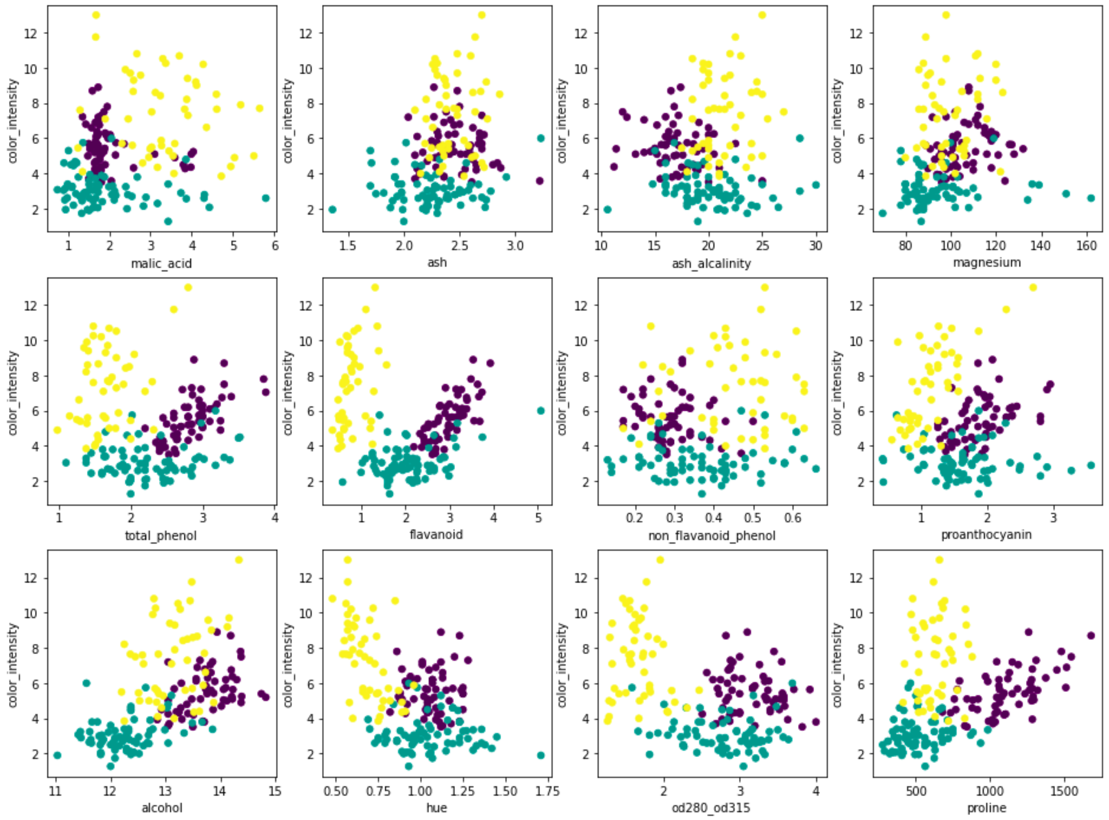
The dataset that has been used here is the wine dataset. The link for the dataset can be found below,
https://archive.ics.uci.edu/ml/datasets/wine

The dataset has 178 instances of 13 different attributes. All the instances are of three classes of wines. All the attributes are real/integer values and there are no missing values in the dataset.

a) To visualize the dimensions of the dataset, scatter plots were plotted. Initially, all the attributes were plotted on the y-axis one by one with the ‘alcohol’ attribute on the x-axis. The scatter plots can be seen in the figure below,

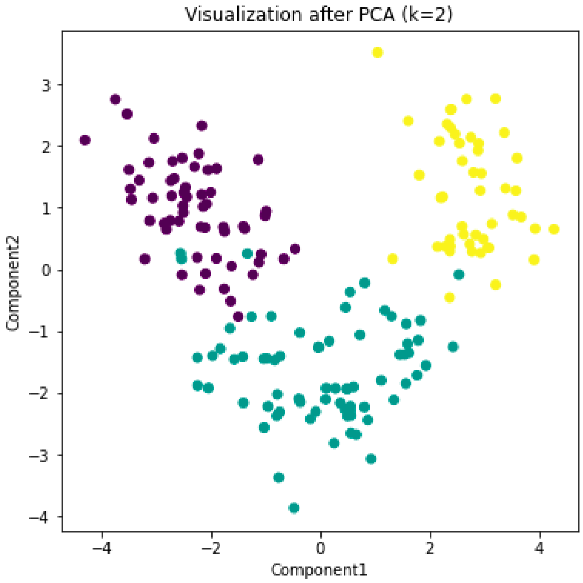


As we can see, the amount of overlap is very high in these cases.
Let’s try with a different attribute. Now, let’s keep the ‘color_intensity’ attribute on the y-axis and analyze it with the other attributes.



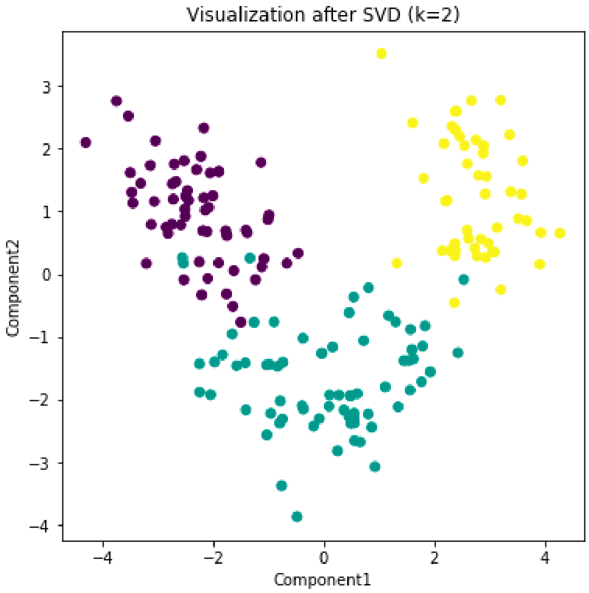
As we can see in the above scatter plots, the amount of overlap is less than the previous cases but we still cannot cluster them properly.
This shows that we can apply PCA and SVD on the dataset for better clarity on the clusters and much less overlap.

b,c) After implementing the PCA logic, the 2-D graph that we obtain when the 2 most important principal components are used is shown below,



The code can be found in google colab (link attached in the beginning).
Here we can see that there is much less overlap after applying PCA on the dataset and the 3 clusters are clearly visible with very less error.

Alternatively, the graph of the 2 best principal components when SVD was applied is shown below,



Here also we can clearly make out the 3 clusters after applying SVD. Infact, this is the same result we got after doing PCA.
Hence, we can say that **k=2 (dimensions)** is sufficient for proper clustering of the dataset when PCA and SVD is used.

d)

- We saw that using PCA and SVD helped us reduce the redundancy in the dataset and minimized the number of dimensions using their logics.
- Using PCA/SVD can in turn improve the performance of the ML algorithm and also improves the visualization.
- With the help of PCA and SVD, we can remove the correlated features.

THANK YOU