IT350 Assignment-1

Name: Durga Supriya HL
Roll no.: 201IT121

Dataset: here
Code: here

1. Download a dataset of your choice

Its dataset related to iris flower and its features

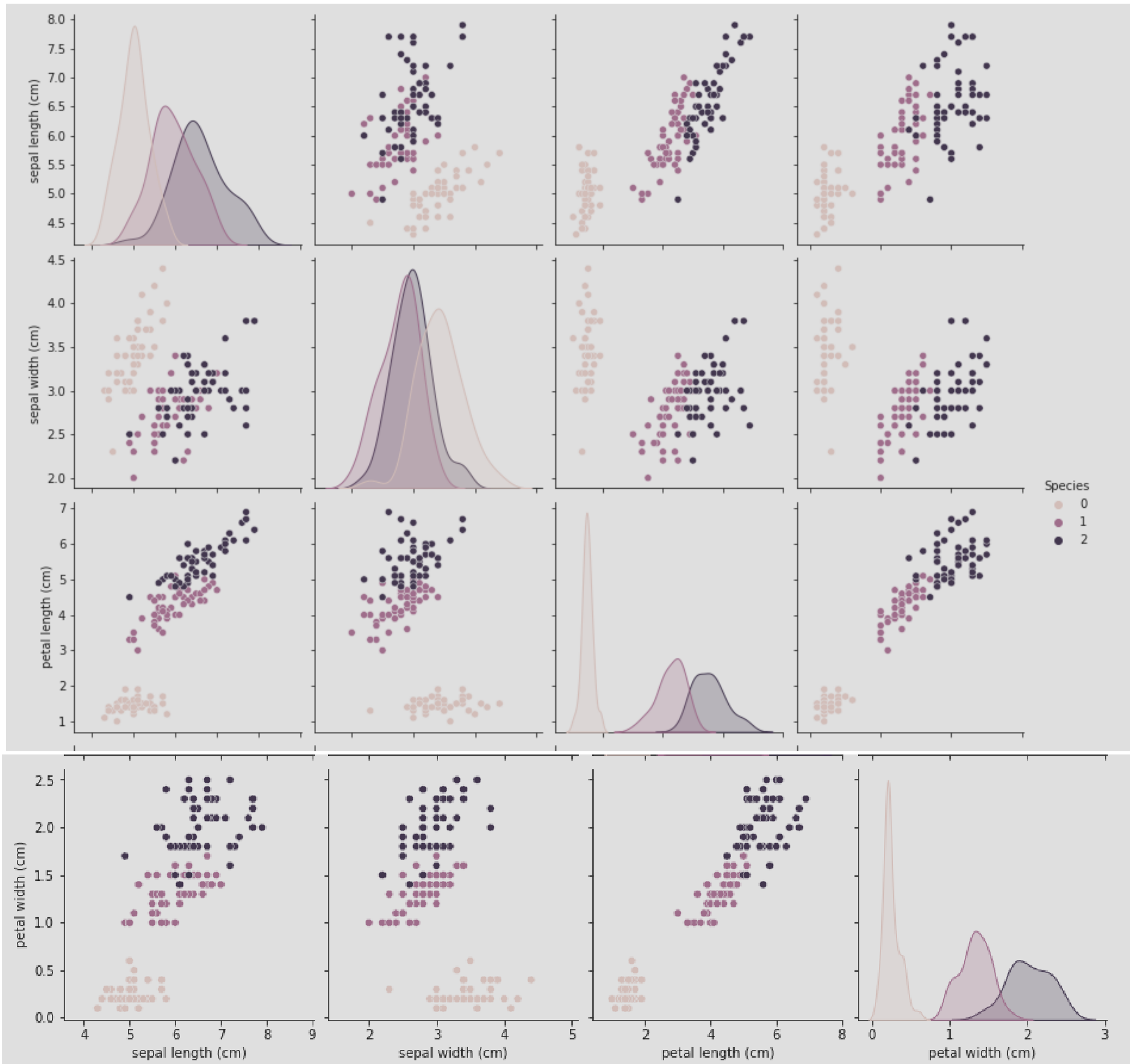
	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	Species
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	2
146	6.3	2.5	5.0	1.9	2
147	6.5	3.0	5.2	2.0	2
148	6.2	3.4	5.4	2.3	2
149	5.9	3.0	5.1	1.8	2

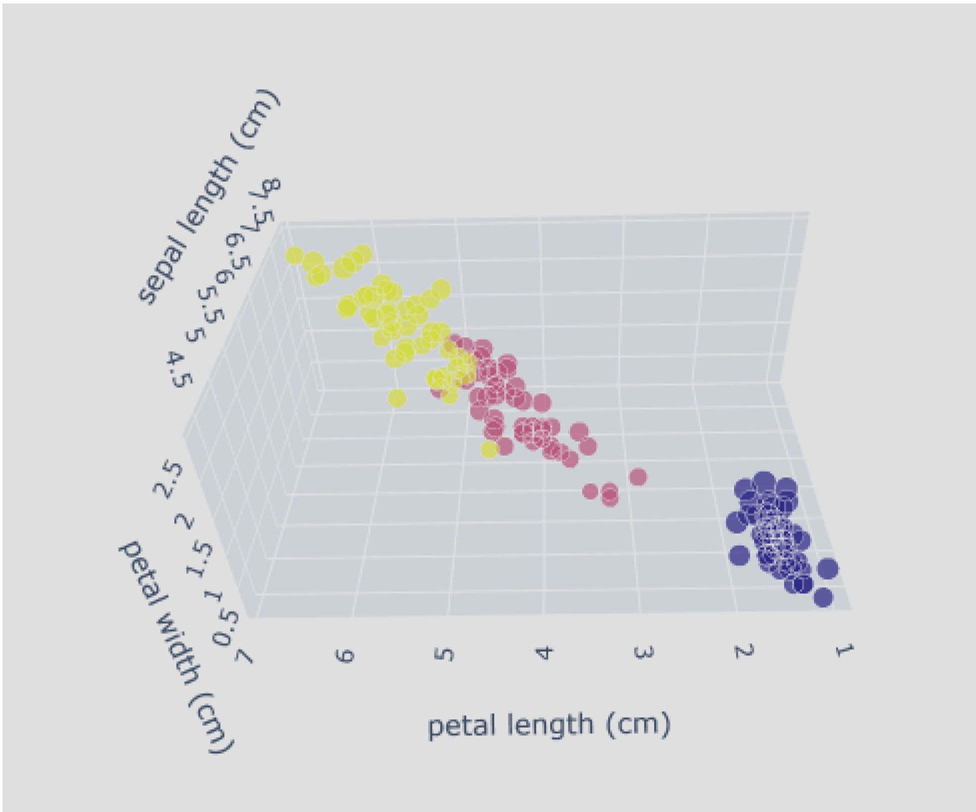
150 rows × 5 columns

2. Visualize it using multiple dimensions and say why SVD and PCA should be used here (2)

The above dataset has 4 features with 150 rows.

Visualisation of the dataset





Singular Value Decomposition(SVD): Using SVD, we can determine the matrix's rank, quantify the linear system's sensitivity to numerical error, or obtain an optimal lower-rank approximation to the matrix. So SVD may help in
reduce the storage space needed

- speed up computation, fewer dimensions mean less computing, also fewer dimensions can allow usage of algorithms unfit for a large number of dimensions
- remove redundant features,
- reducing a data's dimension to 2D or 3D may allow us to plot and visualize it, maybe observe patterns, give us insights

Principal component analysis (PCA) simplifies the complexity of high-dimensional data while retaining trends and patterns. It does this by transforming the data into fewer dimensions, which act as summaries of features.PCA reduces data by geometrically projecting them onto lower dimensions called principal components (PCs), with the goal of finding the best overview of the data using a limited number of PCs.

3. Implement SVD and PCA logic on your own and find the appropriate k-dimensions to represent this data (4)

SVD

Step 1

So, as the first step, we need to find eigenvalues (watch the video provided below to get an understanding of eigenvalues and eigenvectors) of matrix A and as A can be a rectangular matrix, we need to convert it to a square matrix by multiplying A with its transpose.

Step 2

Find square roots of eigenvalues to get singular value matrix and sort them in descending order

Step 3

Sort the vector matrix in the same order as the eigenvalue matrix

Step 4

Remove all the values less than the tolerance value from the singular value matix to get sigma matrix

Step 5

Remove all the values less than the tolerance value from the V matrix

As $A = U\Sigma(V)_T$

$U = AV(\Sigma)_T$

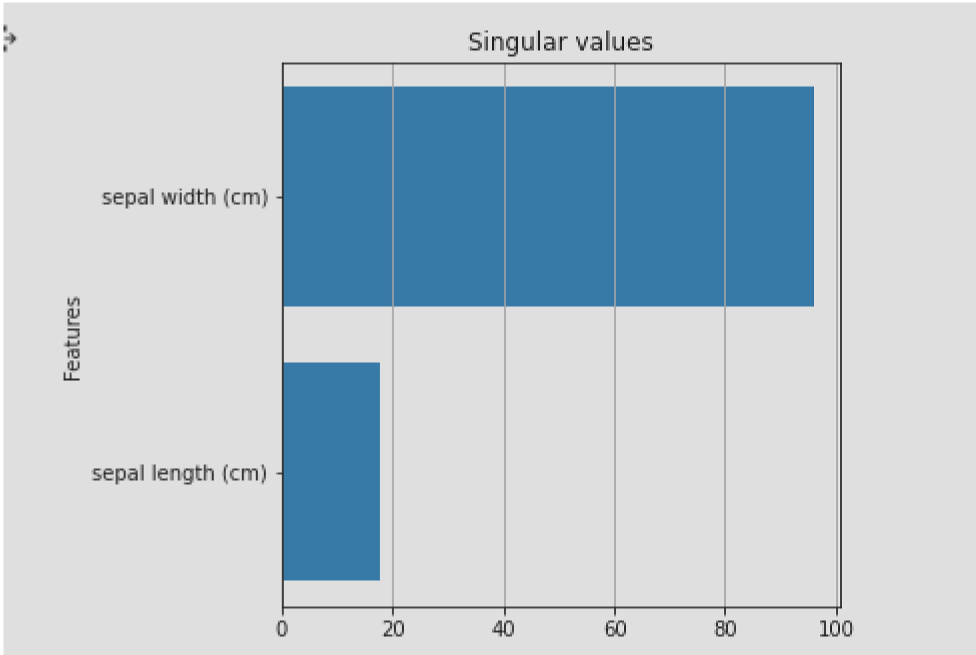
PCA

Principal Component Analysis consists mainly of three steps:

1. First of all, we need to compute the covariance matrix.
2. Once we obtain this matrix, we need to decompose it, using eigendecomposition.
3. Next, we can select the most important eigenvectors based on the eigenvalues, to finally project the original matrix into its reduced dimension.

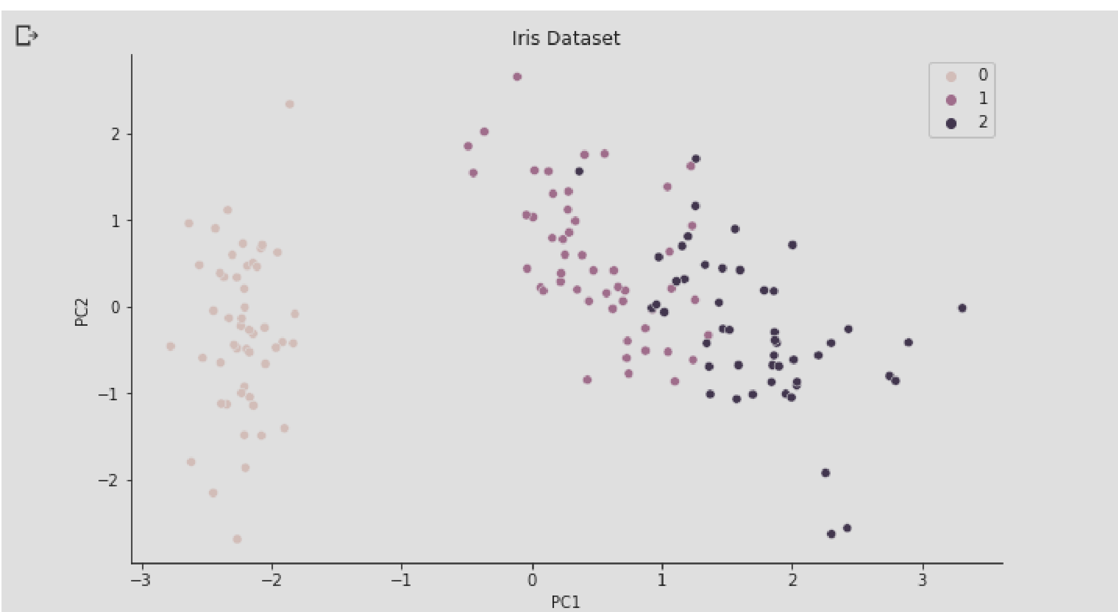
4. Visualize the data after applying SVD and PCA (2

SVD



Reduced it to 2 dimensions

PCA



5. State your conclusions as to how SVD and PCA have helped here. (2)

SVD

Instead of 4 features now we are having only 2 features. This was done based on singular values and the features with singular values less than a particular set tolerance value were removed. This will make further processes in data analysis easier as well as visualization.

PCA

Principal component analysis, or PCA, is a statistical procedure that allows us to summarize the information contained in large data tables by means of a smaller set of “summary indices” that can be more easily visualized and analyzed.