IE 6318 – DATA MINING AND ANALYTICS

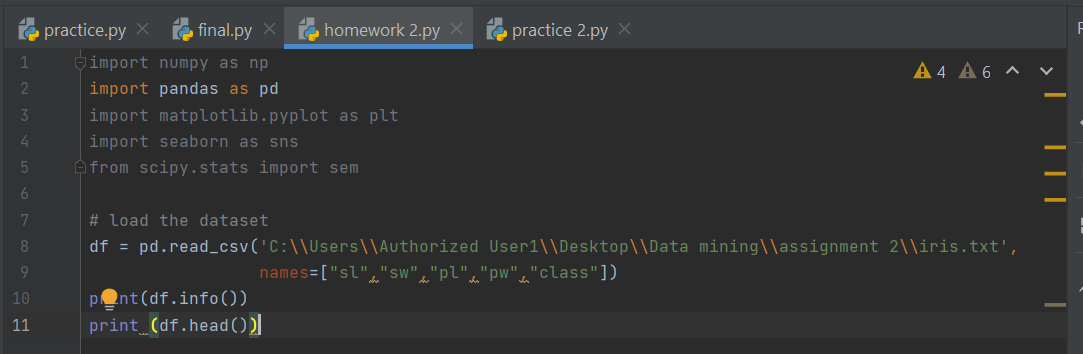
(HW 2)

Jayant Madan

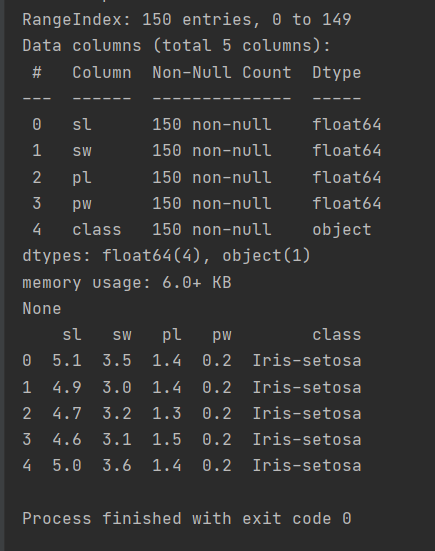
UTA-ID:1001814817

March 14,2021

**1. Explore the Iris dataset and report the following:**

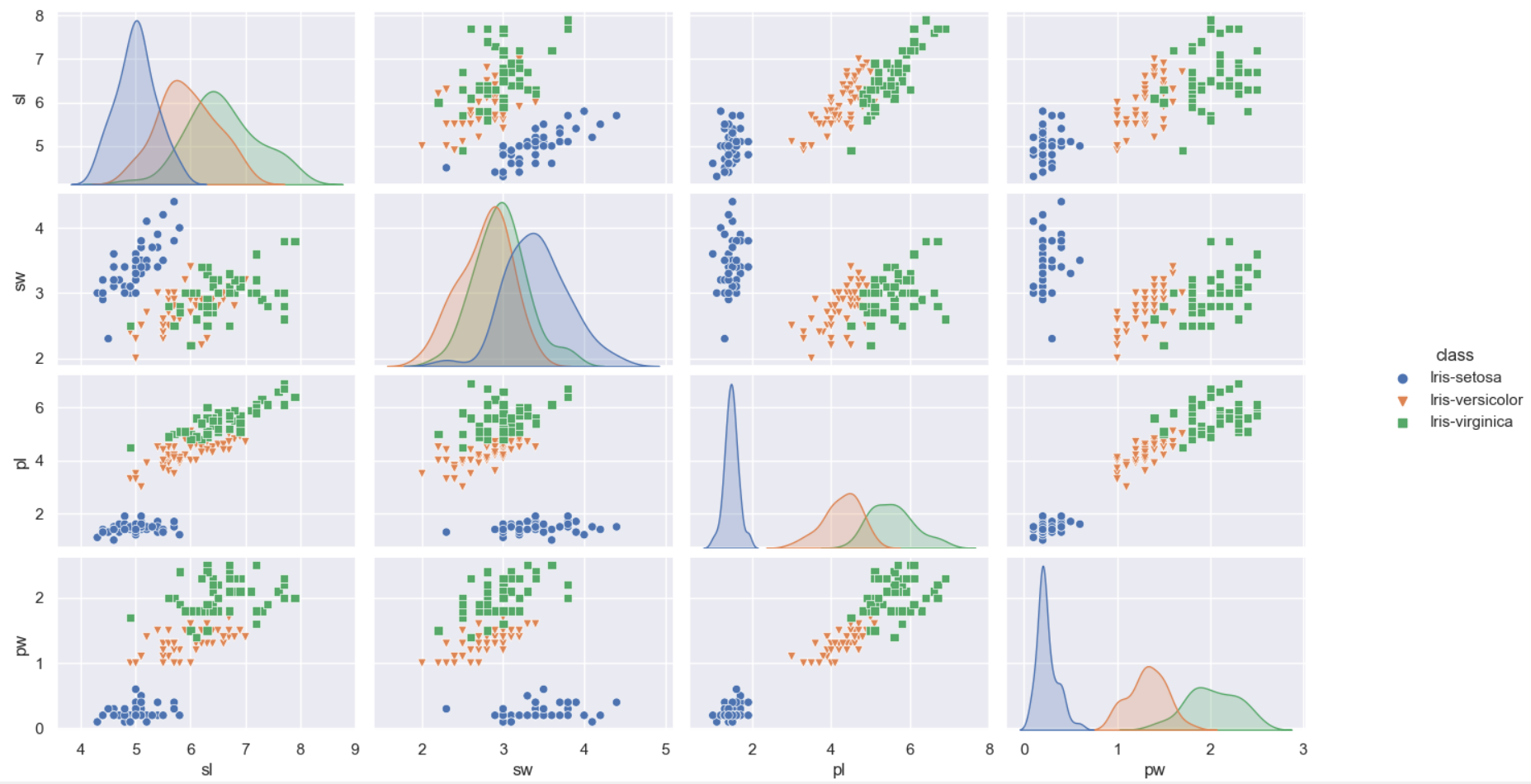
****

sl = sepal length, sw = sepalwidth, pl = petal length, pw = petal width

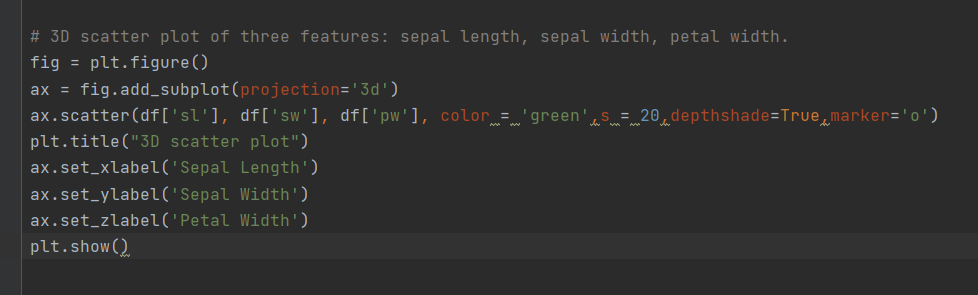
****

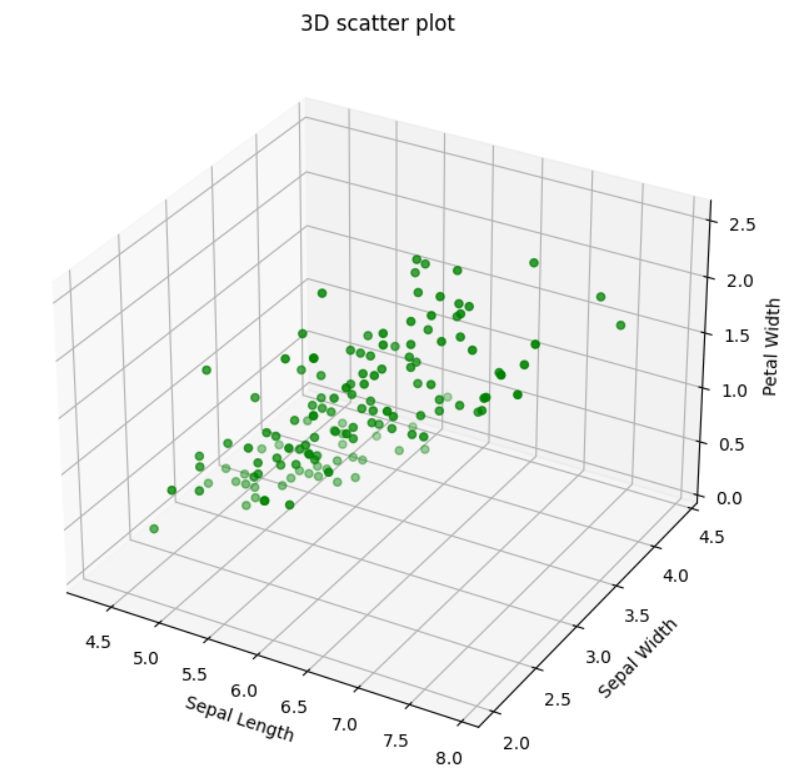
1. 2D scatter plots of the four features (creates a matrix of scatter plots for each pair of the features)



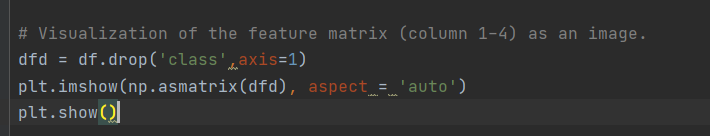


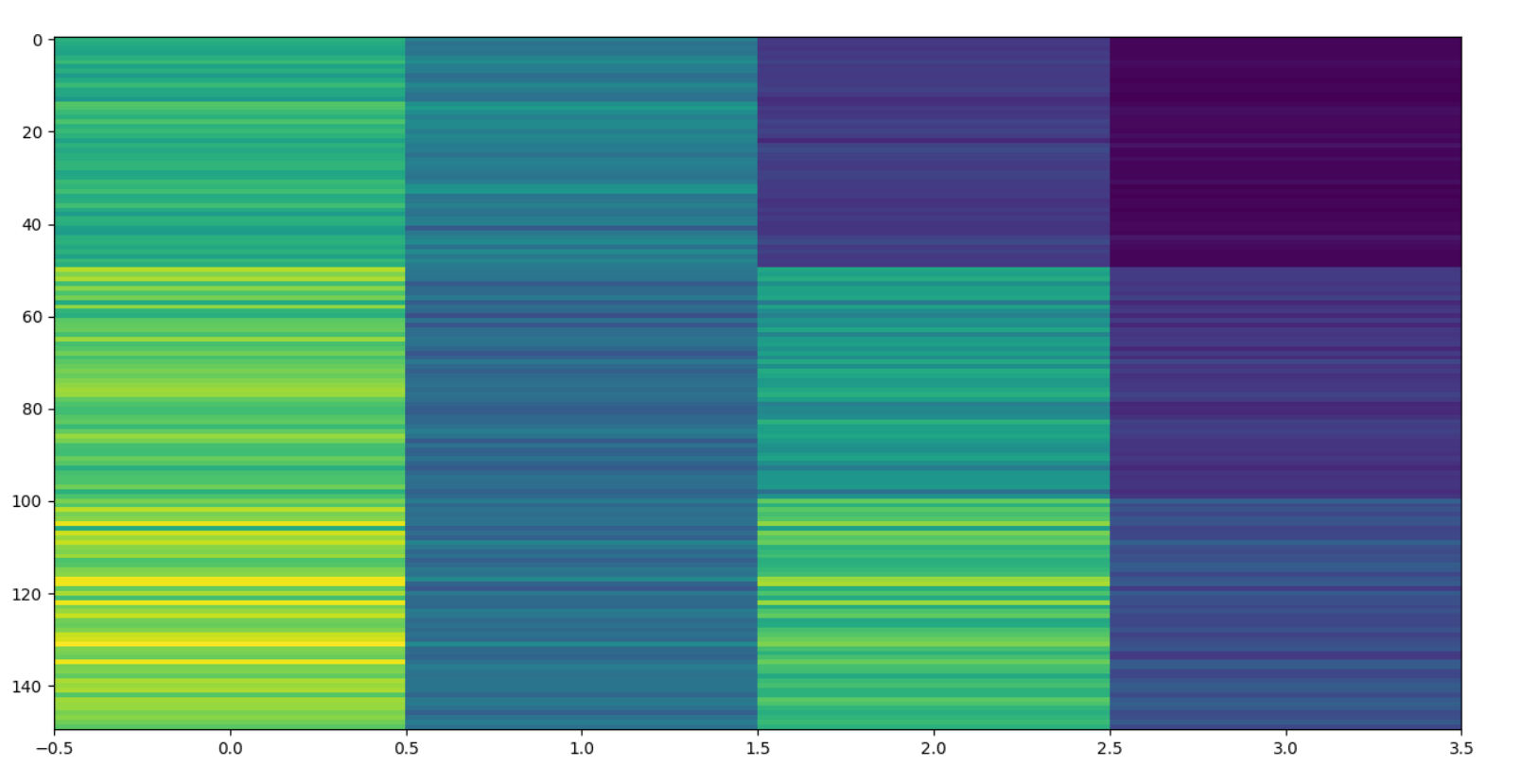
1. 3D scatter plot of three features: sepal length, sepal width, petal width.



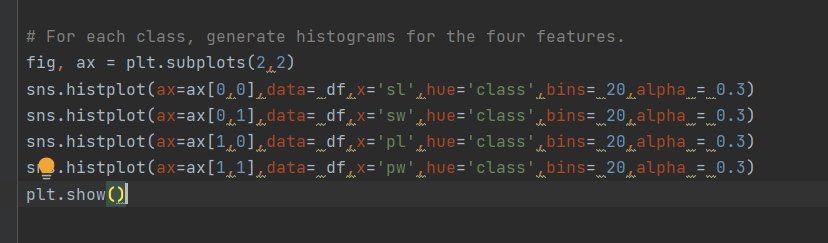


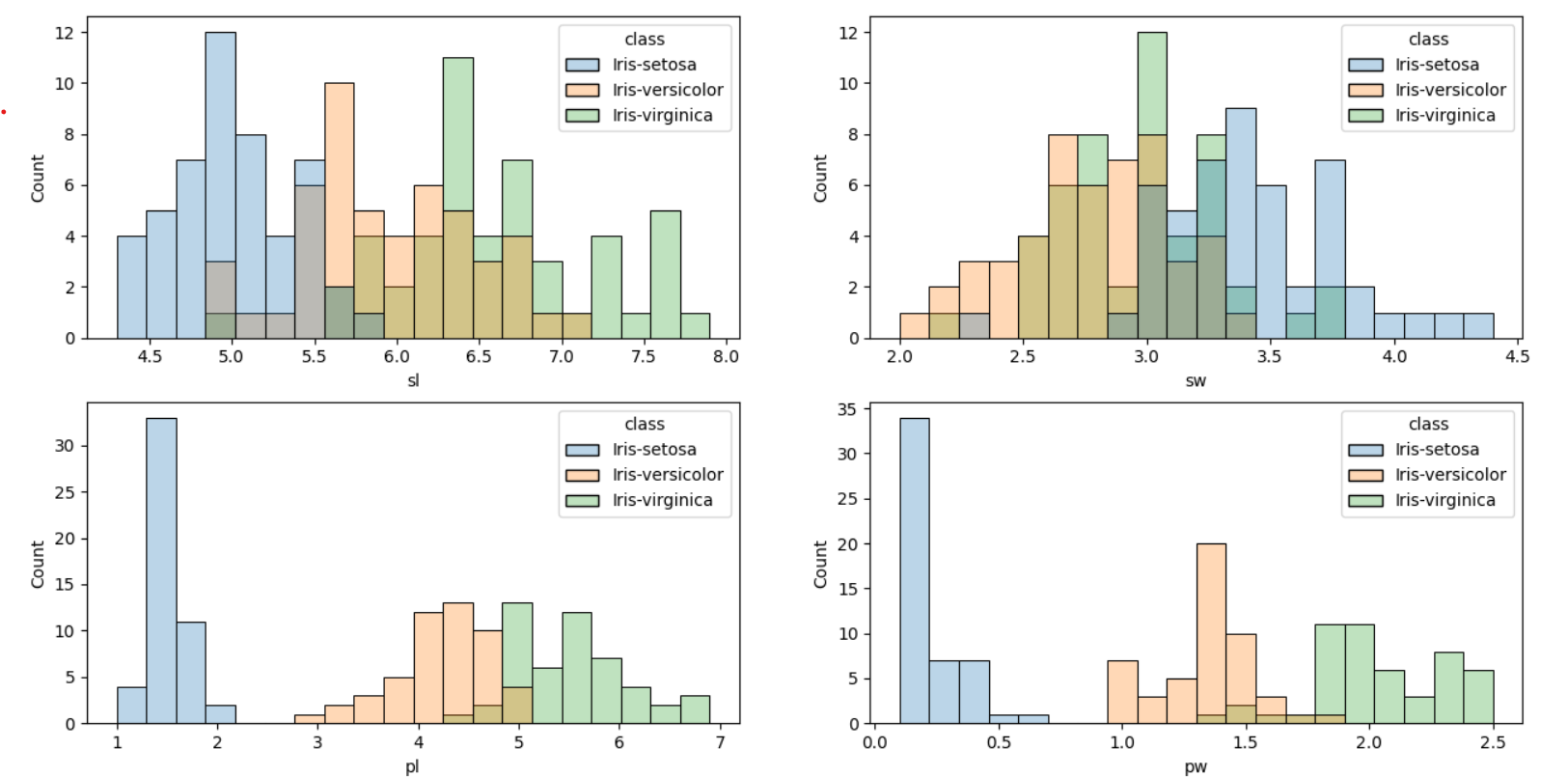
3) Visualization of the feature matrix (column 1-4) as an image.



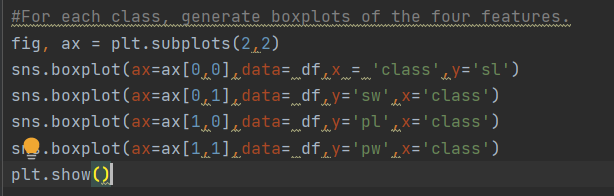


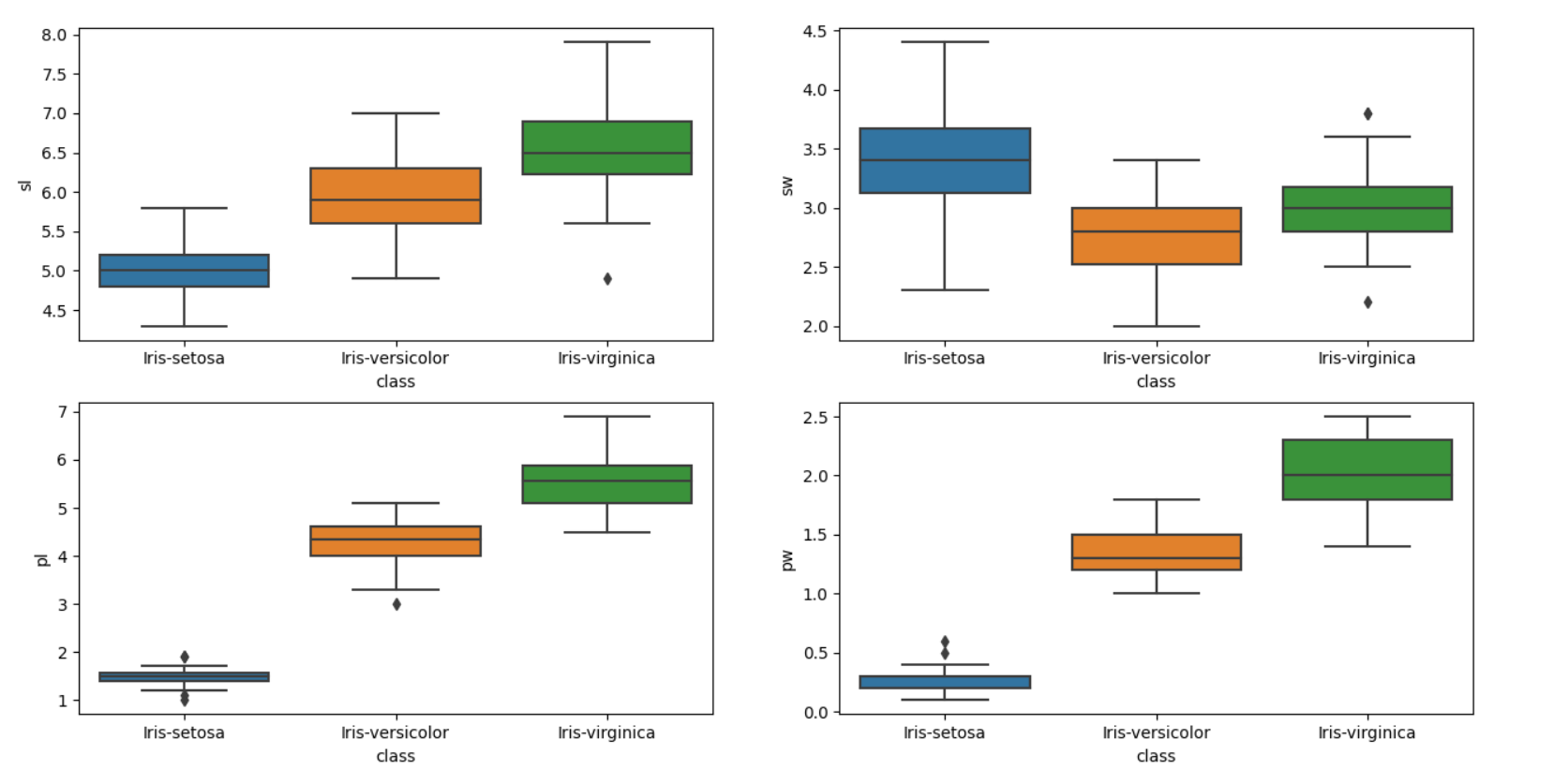
4) For each class, generate histograms for the four features.



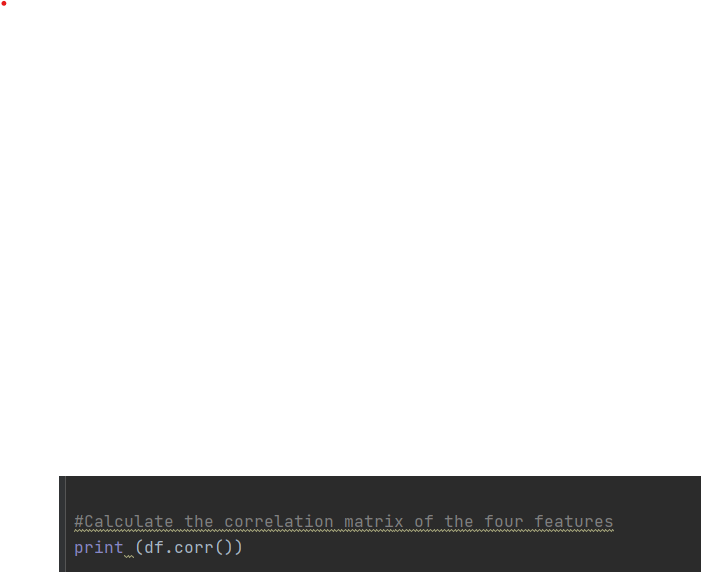


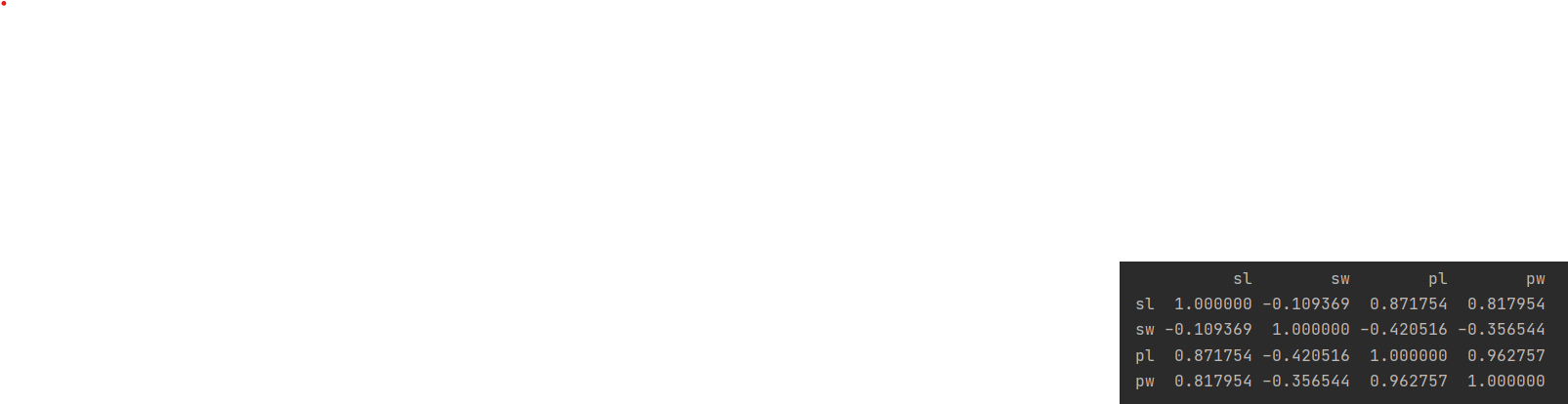
5) For each class, generate boxplots of the four features.



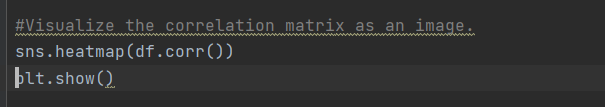


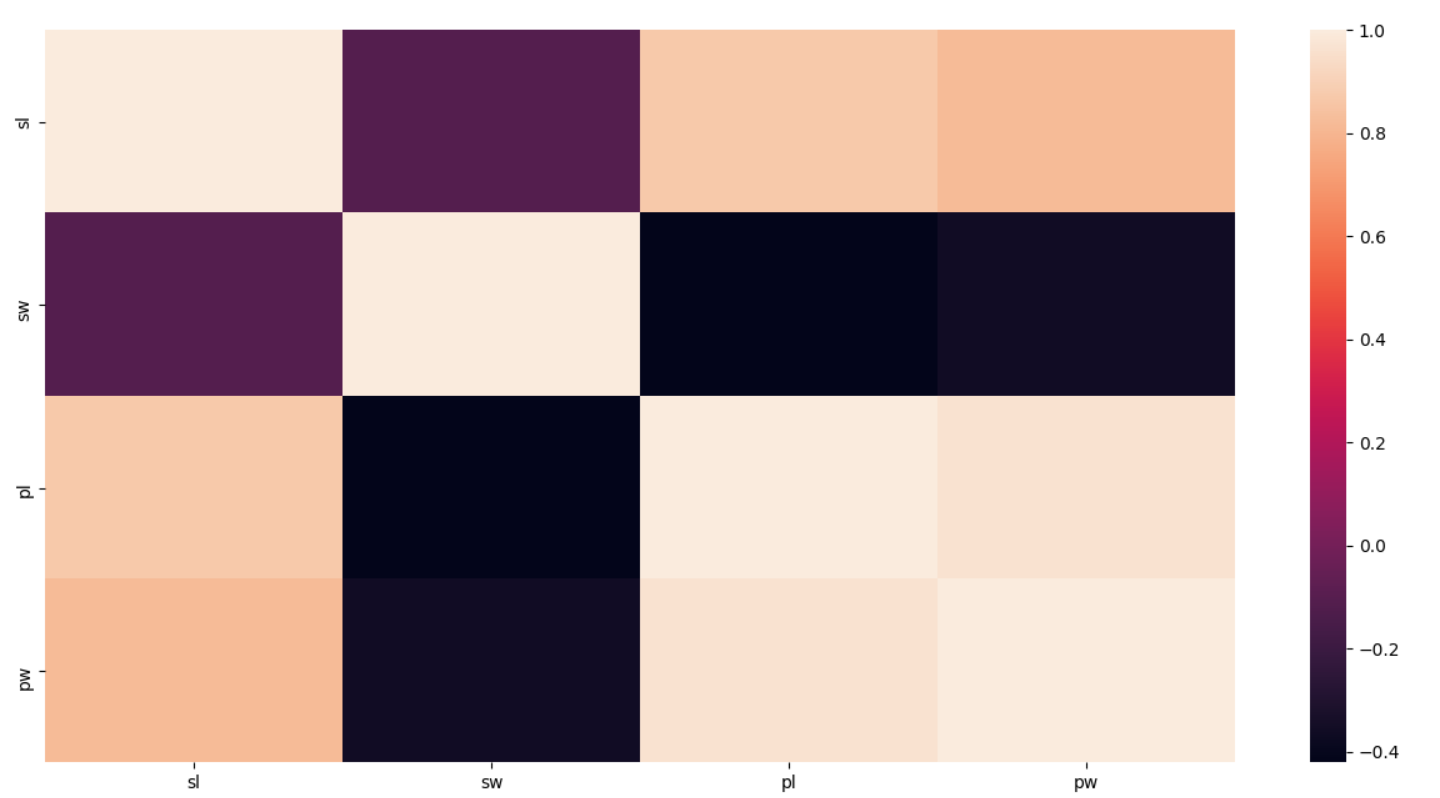
6) Calculate the correlation matrix of the four features



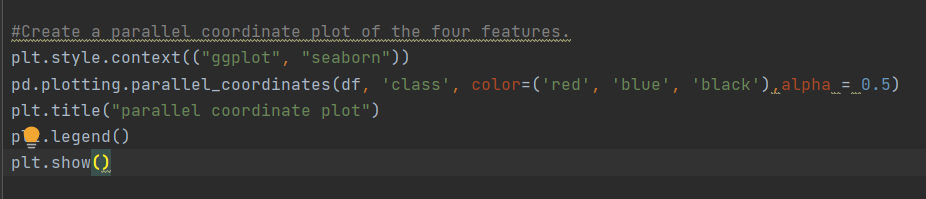


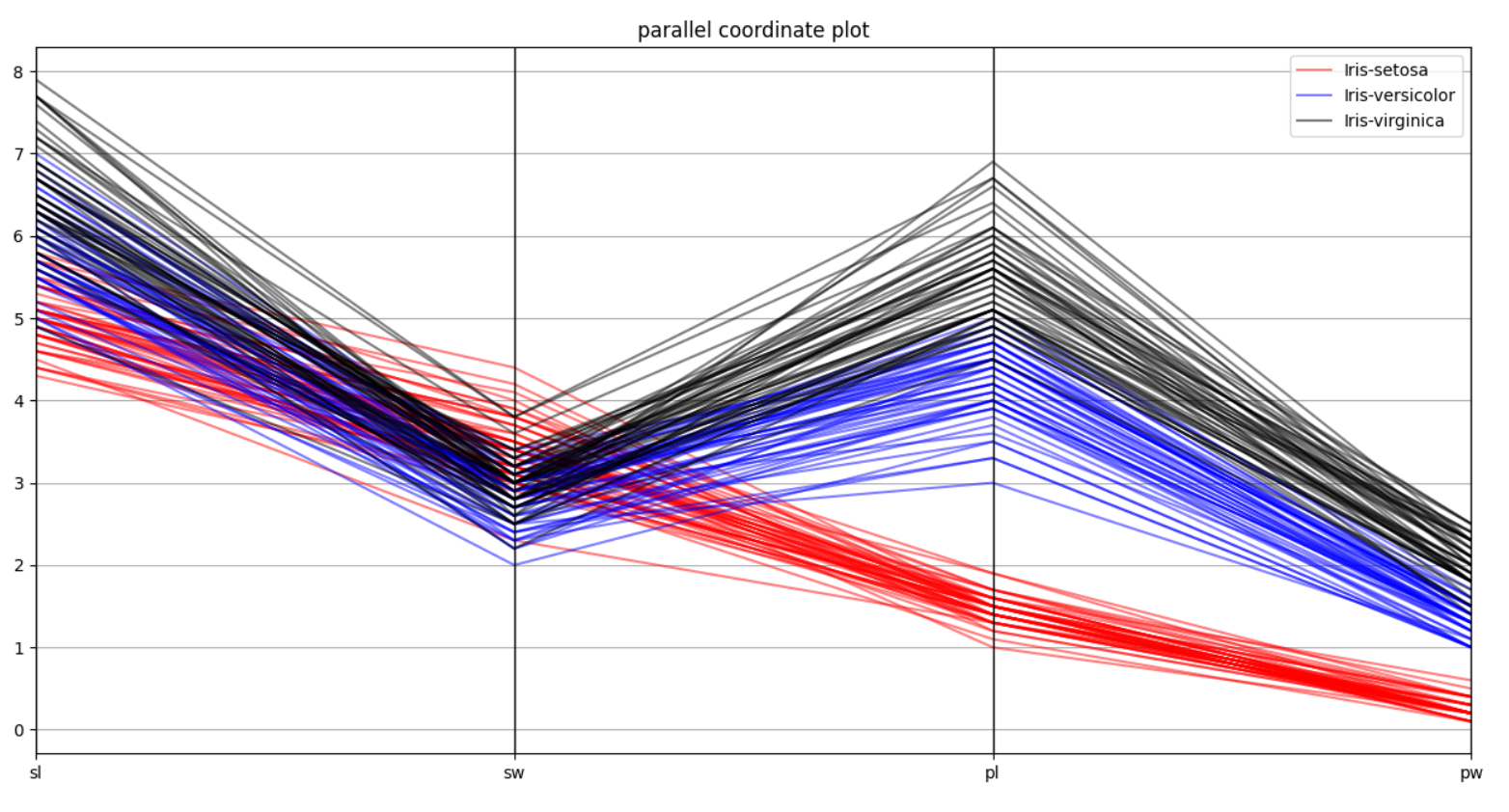
7) Visualize the correlation matrix as an image.





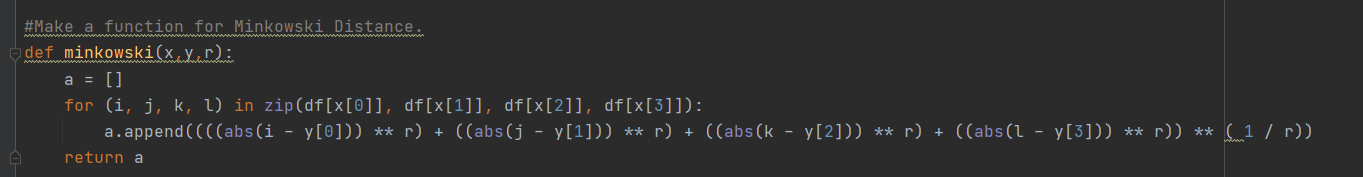
8) Create a parallel coordinate plot of the four features.



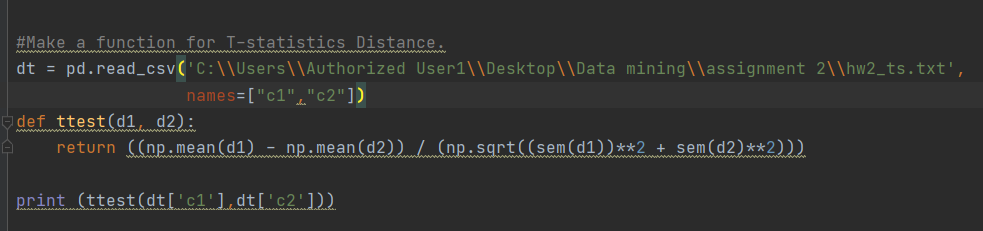


**2. Practice Data Distance Measures**

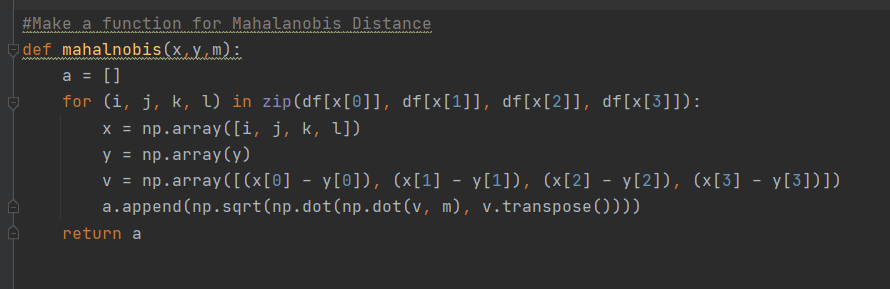
1) Make a function for Minkowski Distance. (3 function inputs: vector A, vector B, and order r)



2) Make a function for T-statistics Distance. (3 function inputs: vector A, vector B)

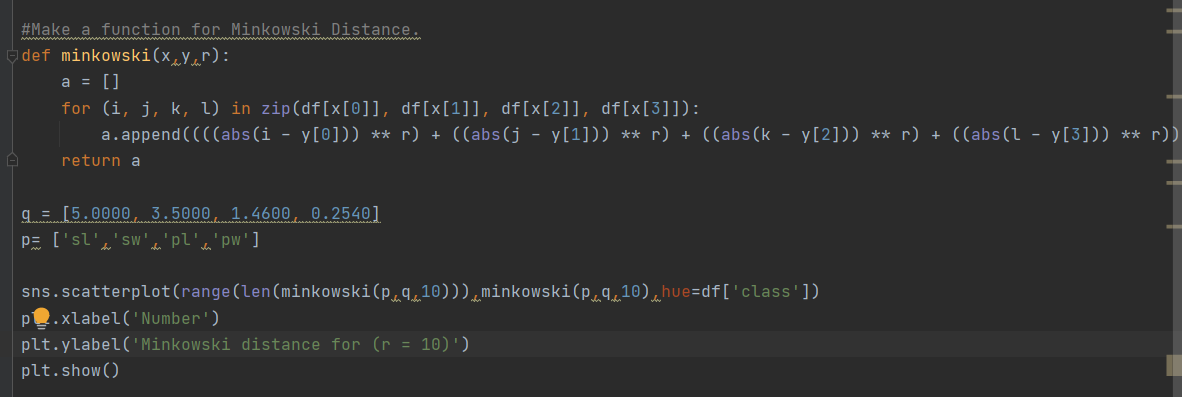


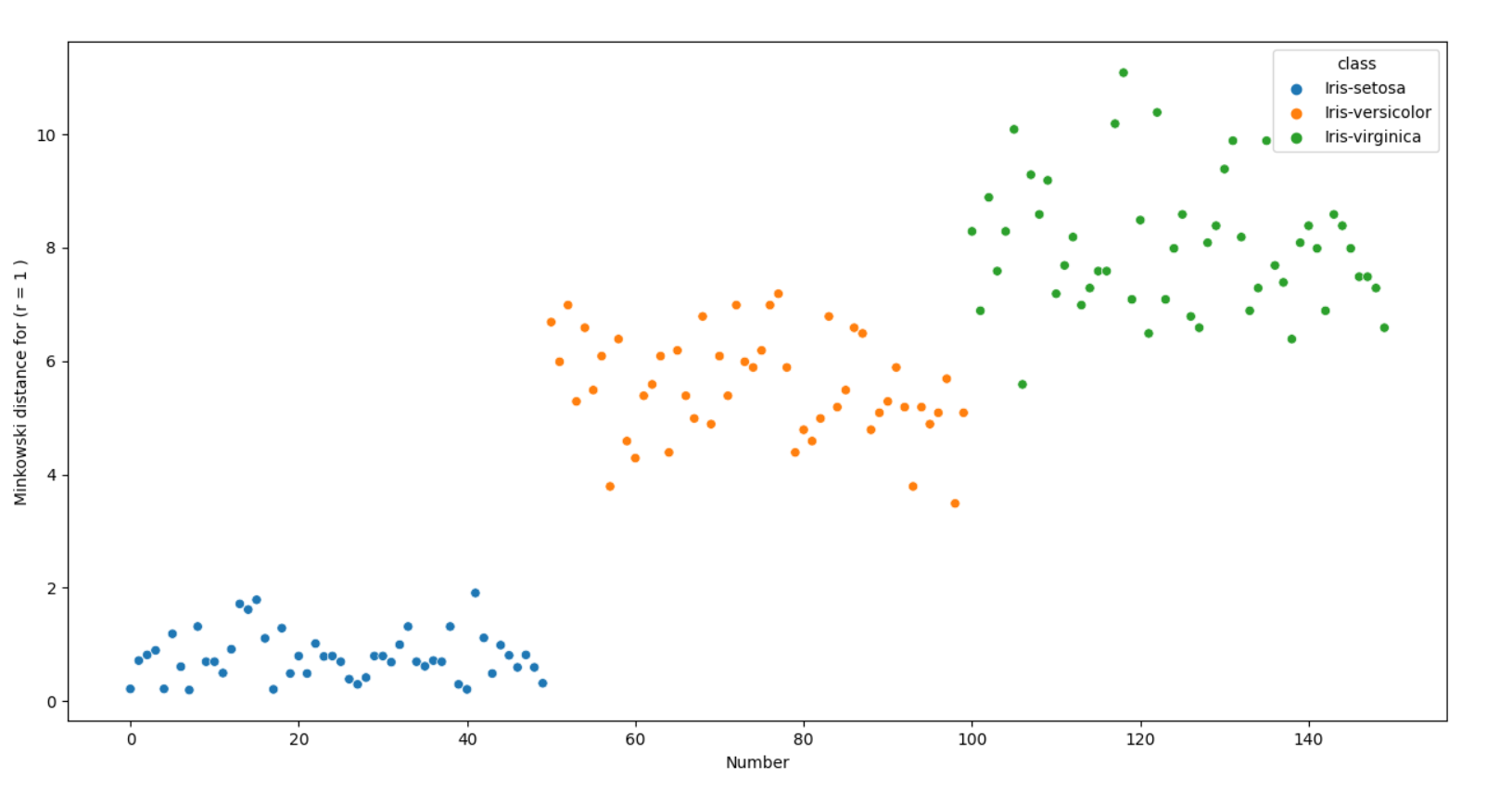
3) Make a function for Mahalanobis Distance. (3 inputs: vector A, vector B, and covariance matrix)

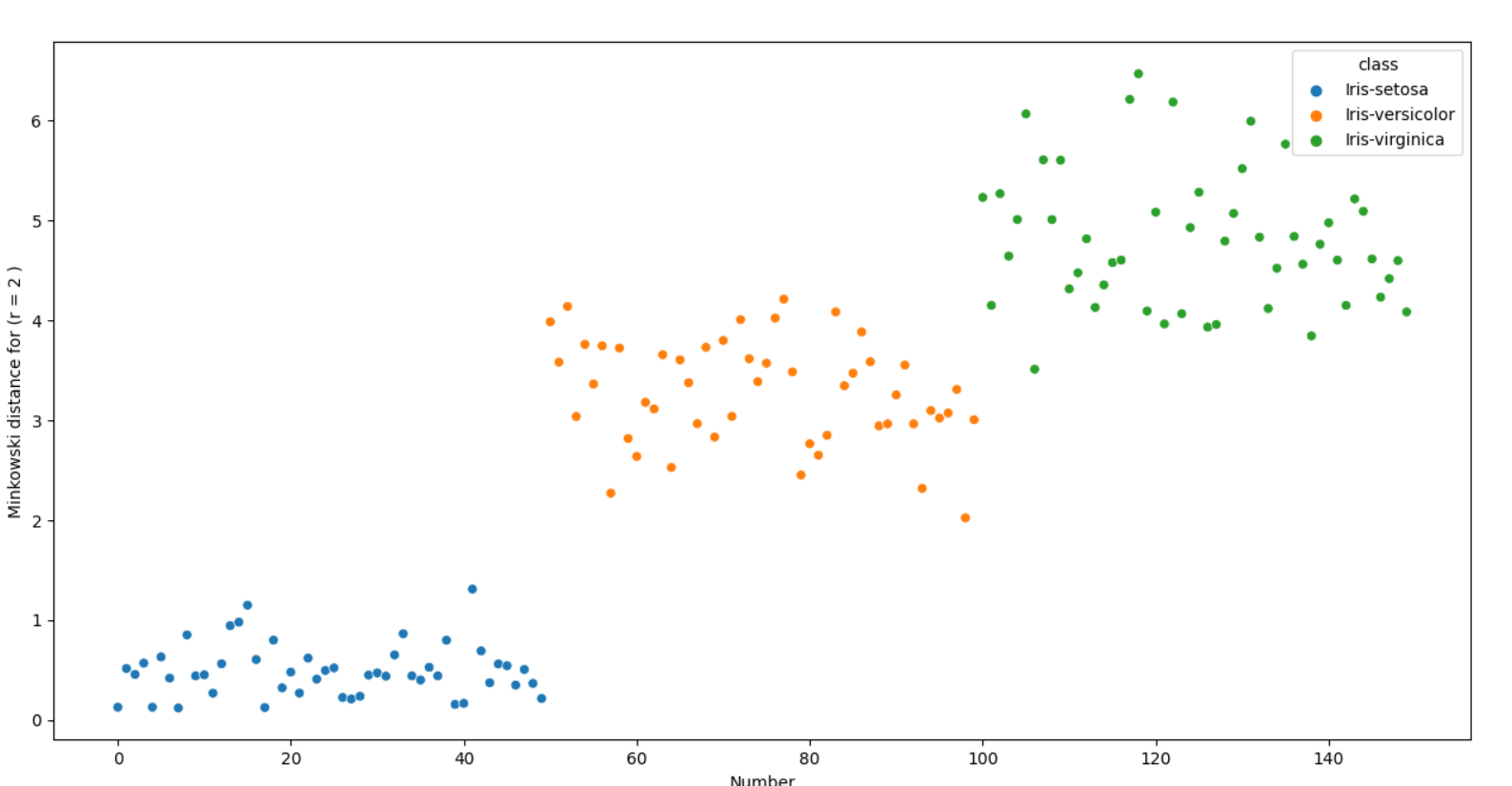


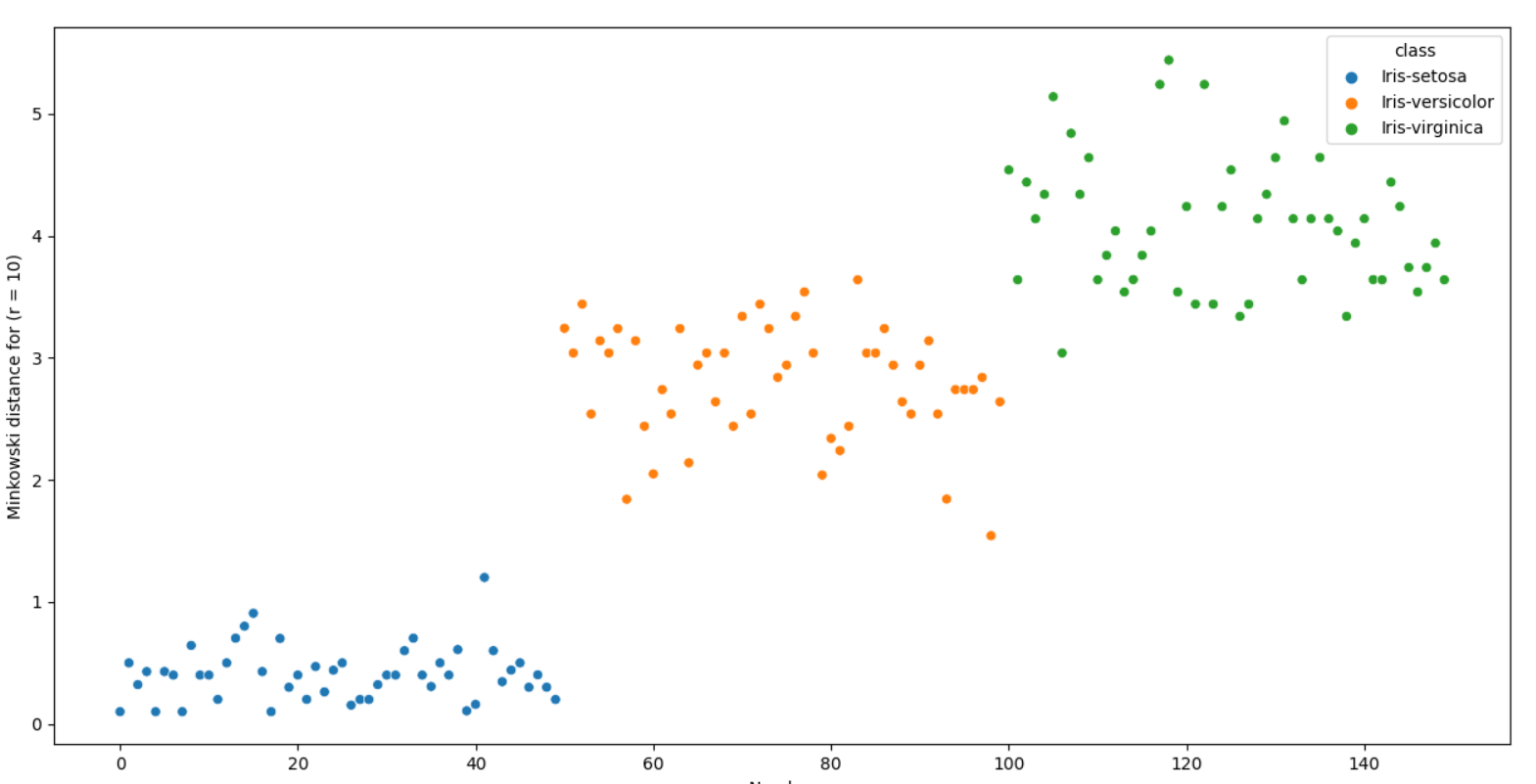
**3. For a new iris data sample S with a feature vector of [5.0000, 3.5000, 1.4600, 0.2540], calculate the distances between the new sample and the 150 samples in the iris dataset using the distance functions you made:**

1) Calculate Minkowski distances with r = 1, 2, 10, respectively, and plot the obtained distances.

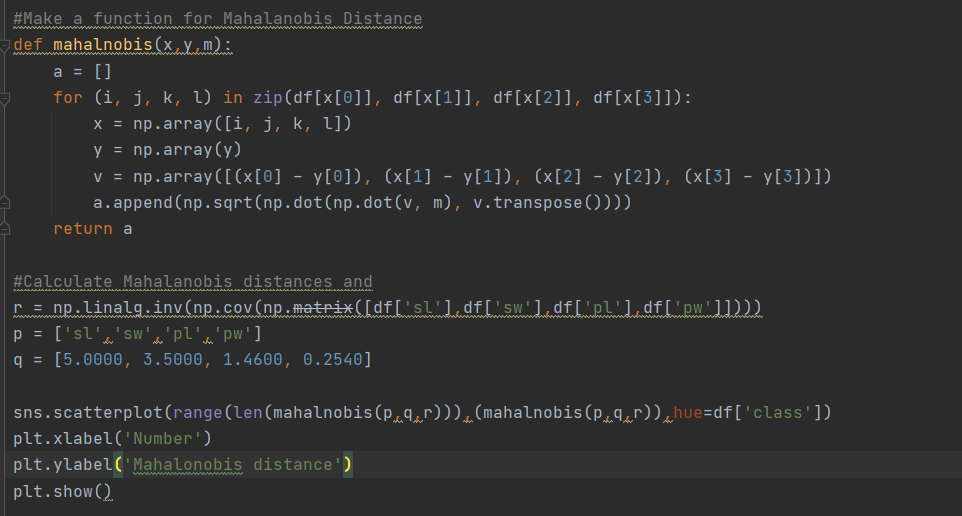


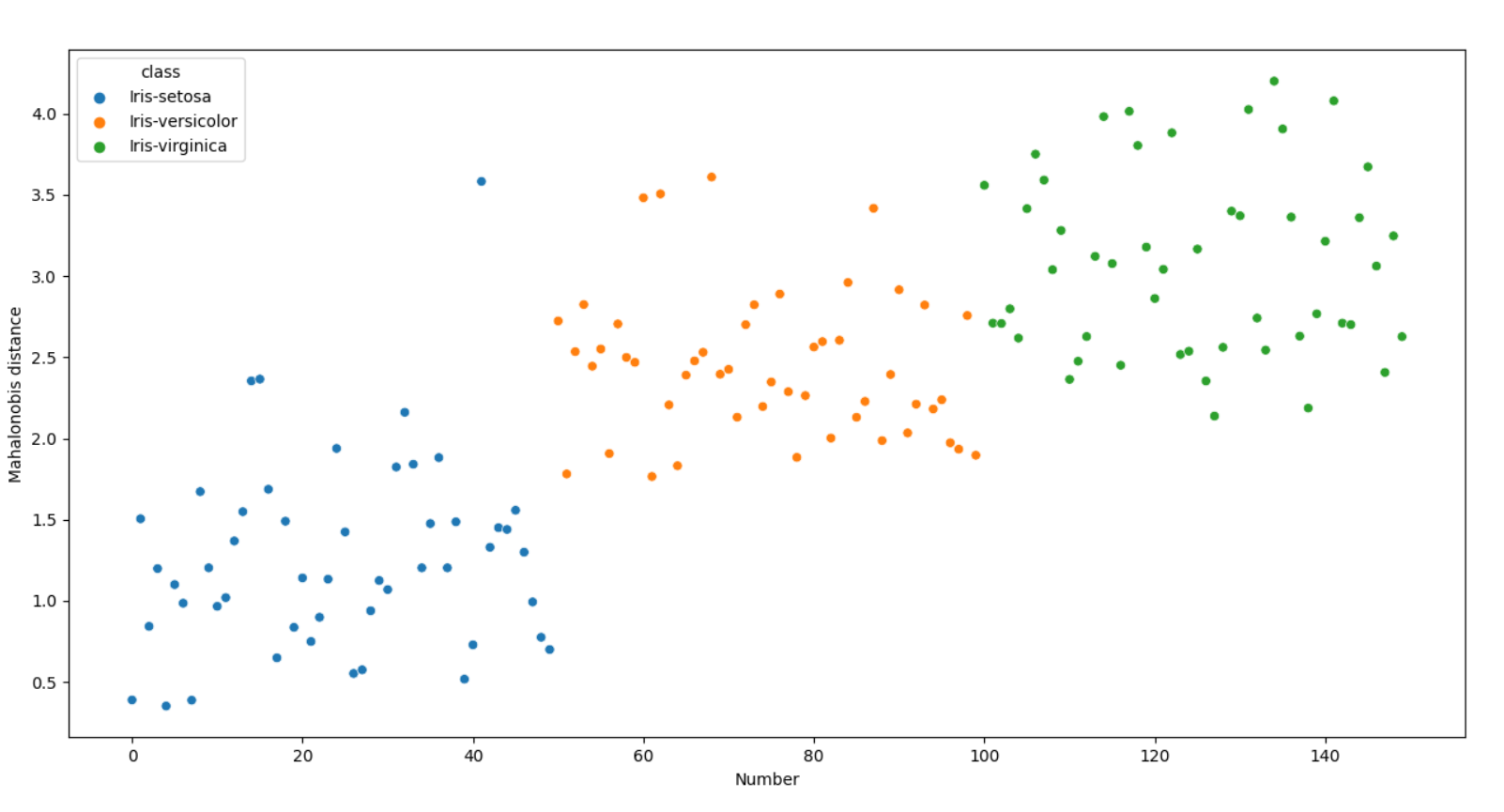






2) Calculate Mahalanobis distances and plot the obtained distances.

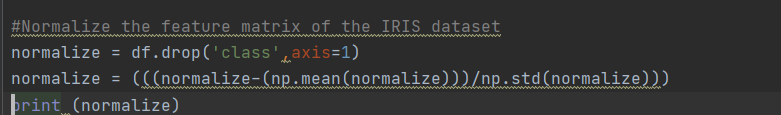


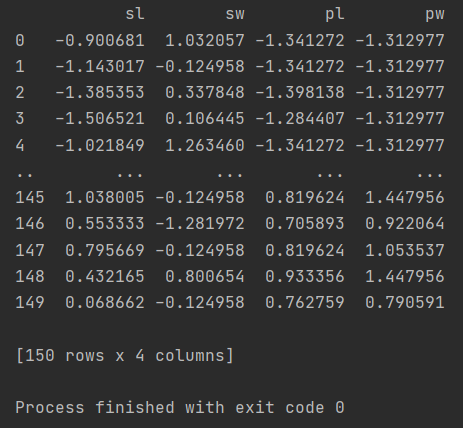


**4. Feature matrix normalization**

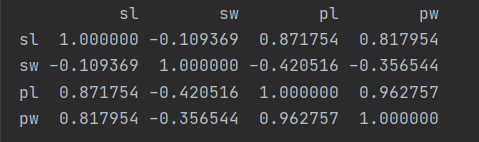
1) Normalize the feature matrix of the IRIS dataset such that each feature has a mean of 0 and a

standard deviation of 1 after normalization.



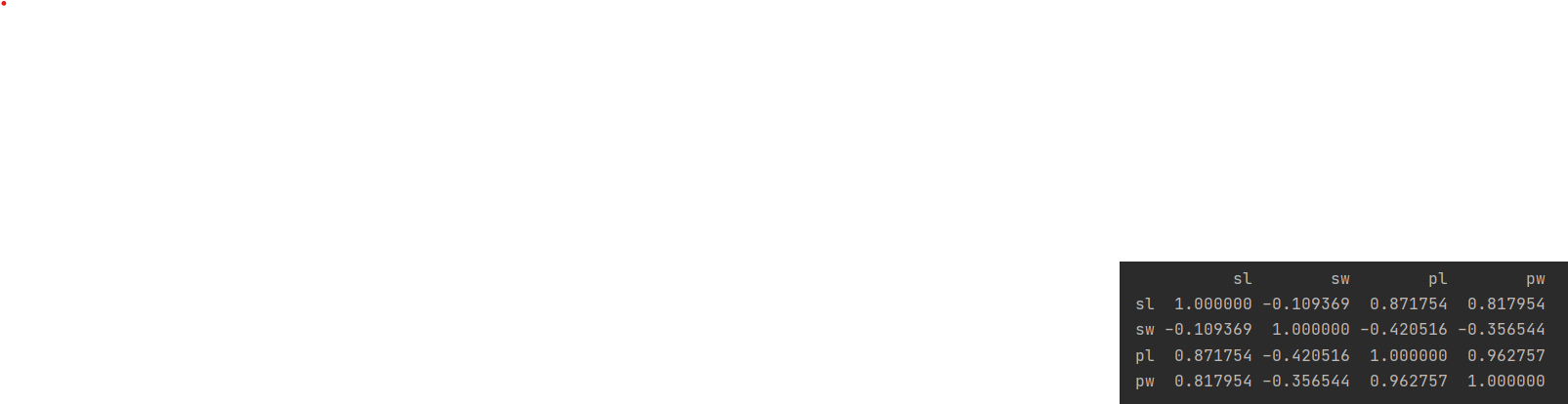
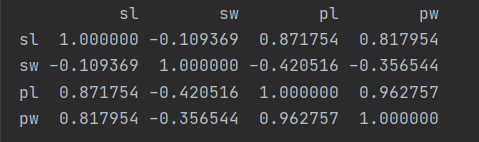


1. Calculate the correlation matrix of the four features after normalization.



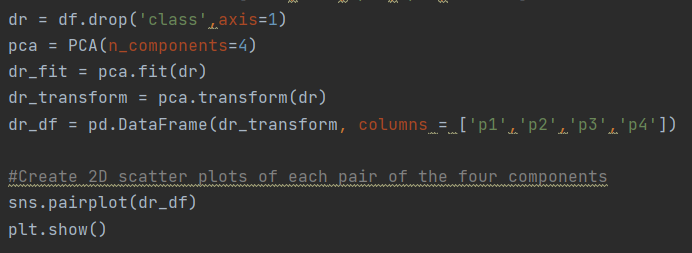
3) Compare the correlation matrix before and after normalization. If they are the same?

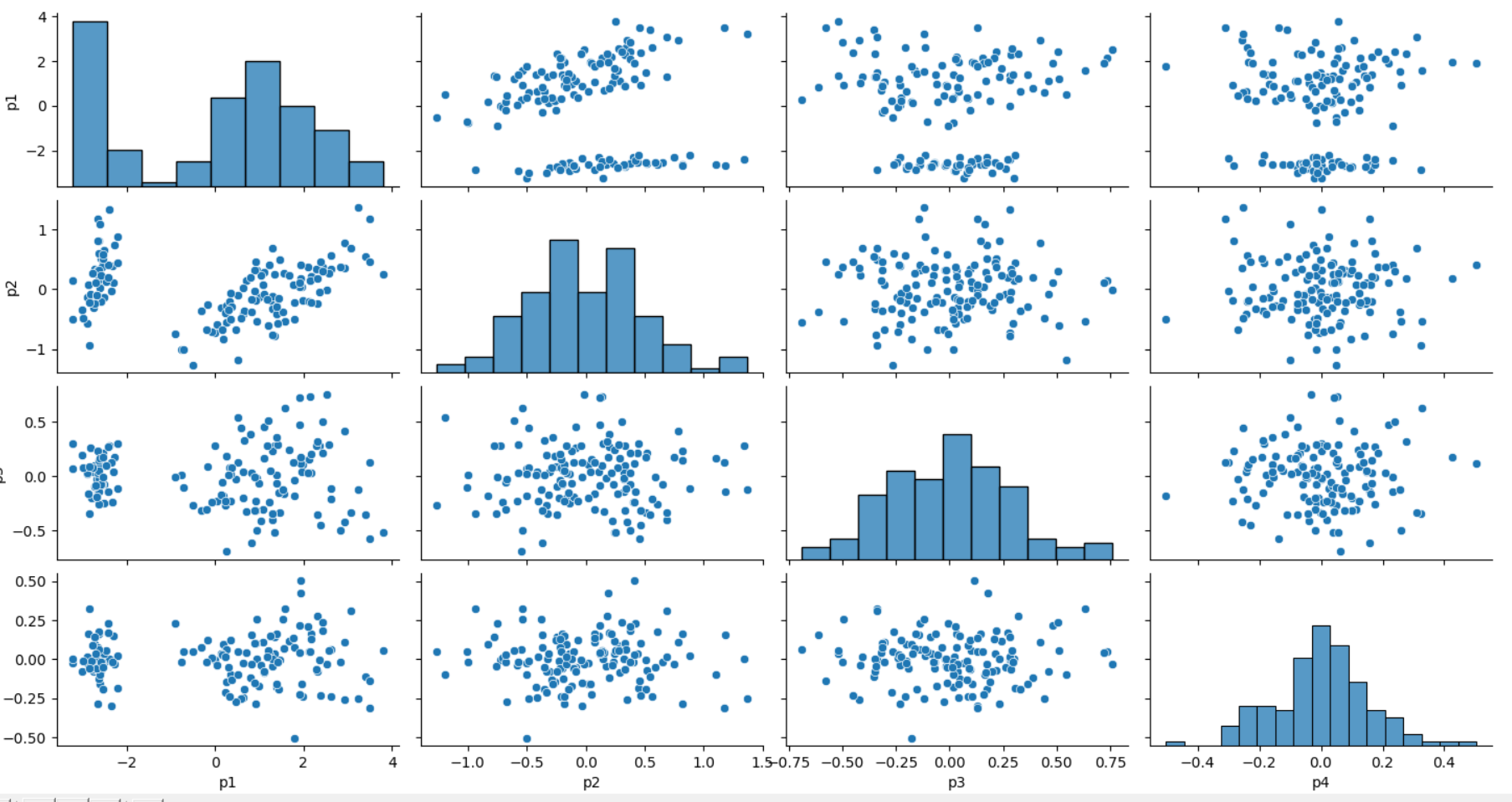
They are same



**5. Principle Component Analysis (PCA) on the IRIS dataset**

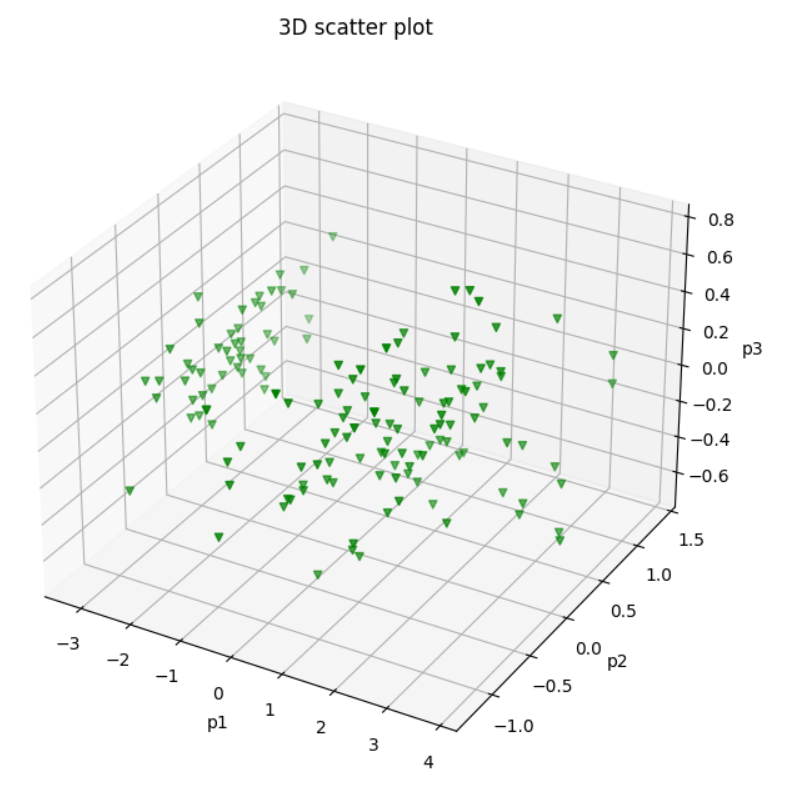
1) Create 2D scatter plots of each pair of the four components



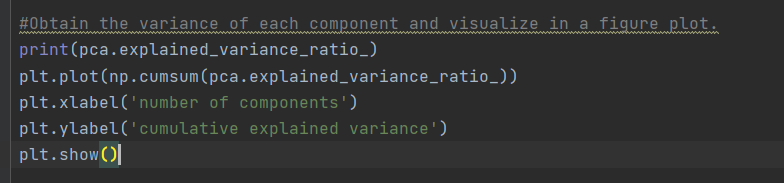


2) 3D scatter plot of the first three components

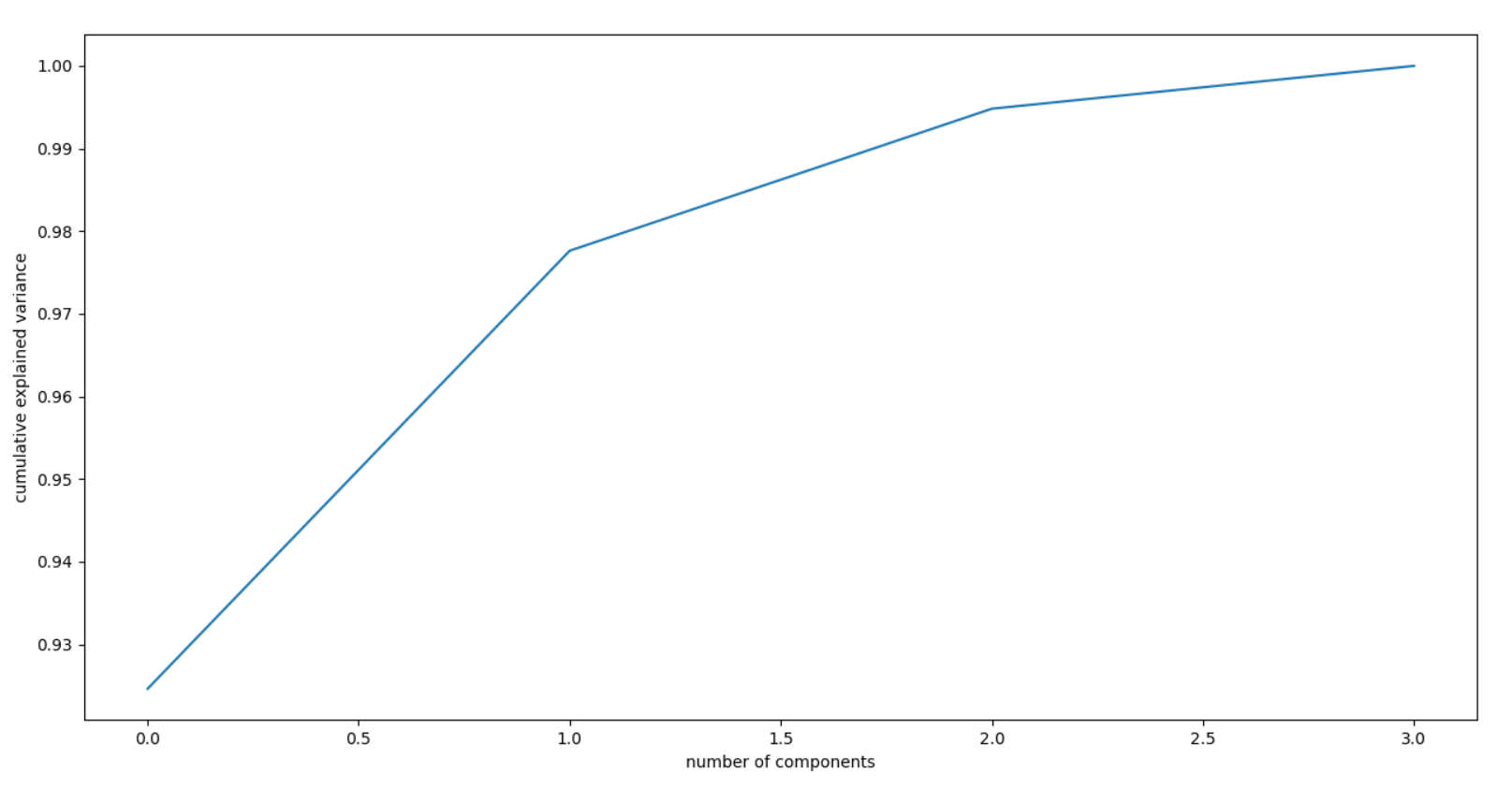




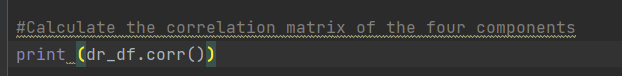
3) Obtain the variance of each component and visualize in a figure plot.

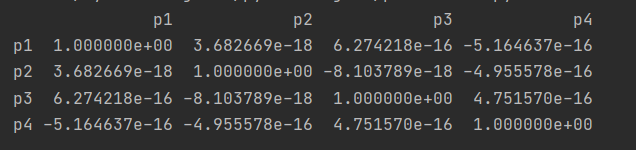






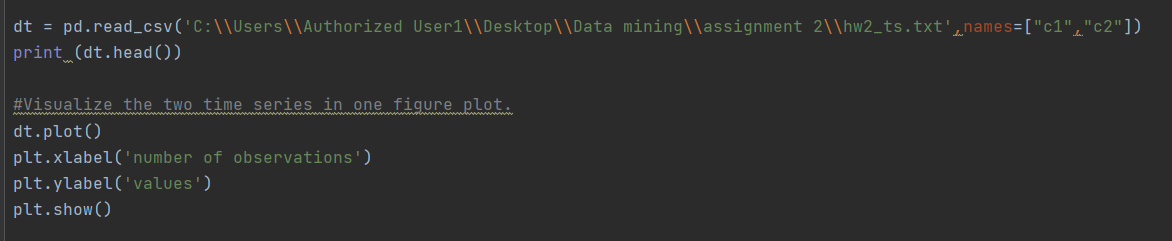
4) Calculate the correlation matrix of the four components

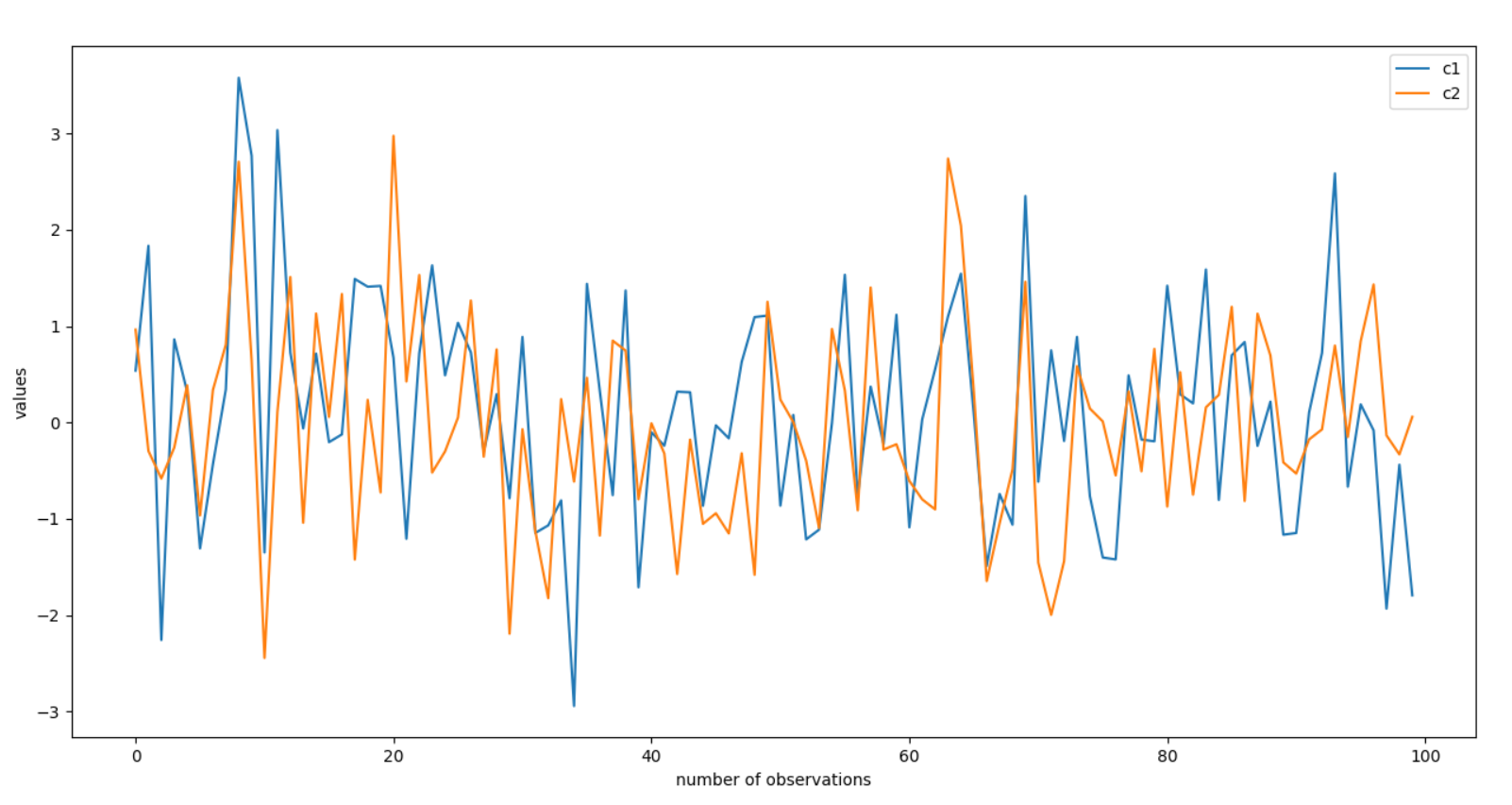




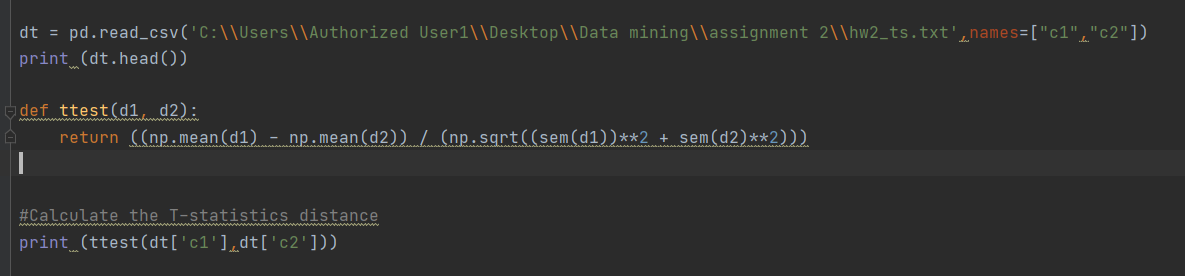
**6. For the dataset with two time series in the “hw2\_ts.txt” file, perform the following analysis:**

1) Visualize the two time series in one figure plot.





2) Calculate the T-statistics distance between the two time series using the function you made.





3) Calculate the correlation of the two time series

