Generic Elective – Computer Science

**Data Analysis and Visualisation using Python**

**Semester – II**

**2023-24**

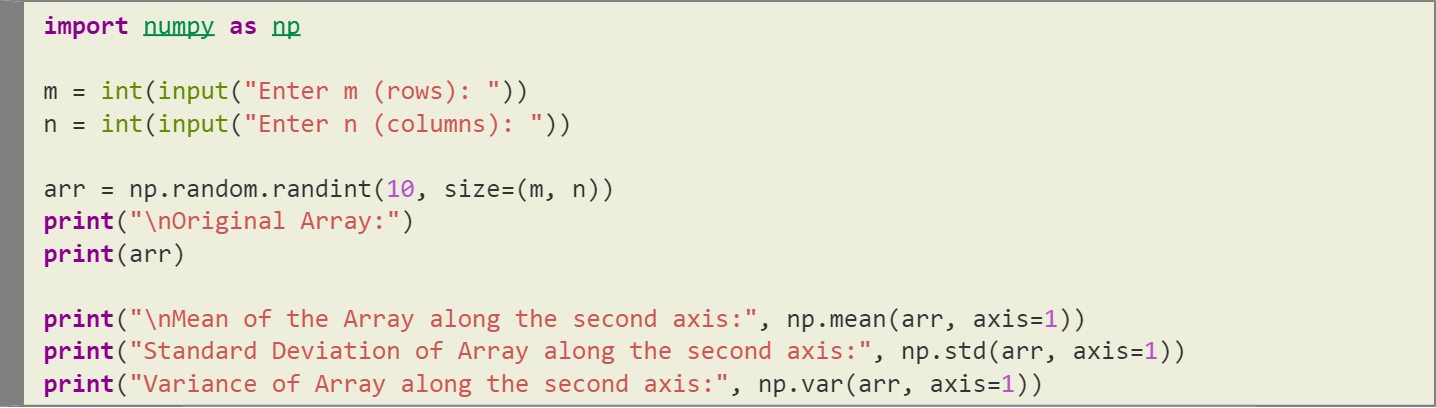
**PRACTICAL FILE**

Lucky Lao

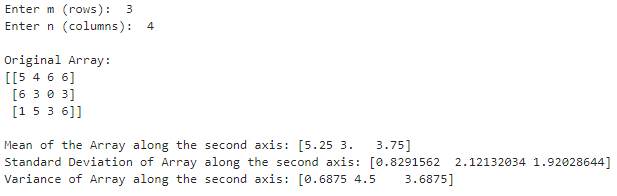
B.Sc. (Hons.) Mathematics

23BMAT019

1. Write programs in Python using NumPy library to do the following:
   1. Compute the mean, standard deviation, and variance of a two dimensional random integer array along the second axis.

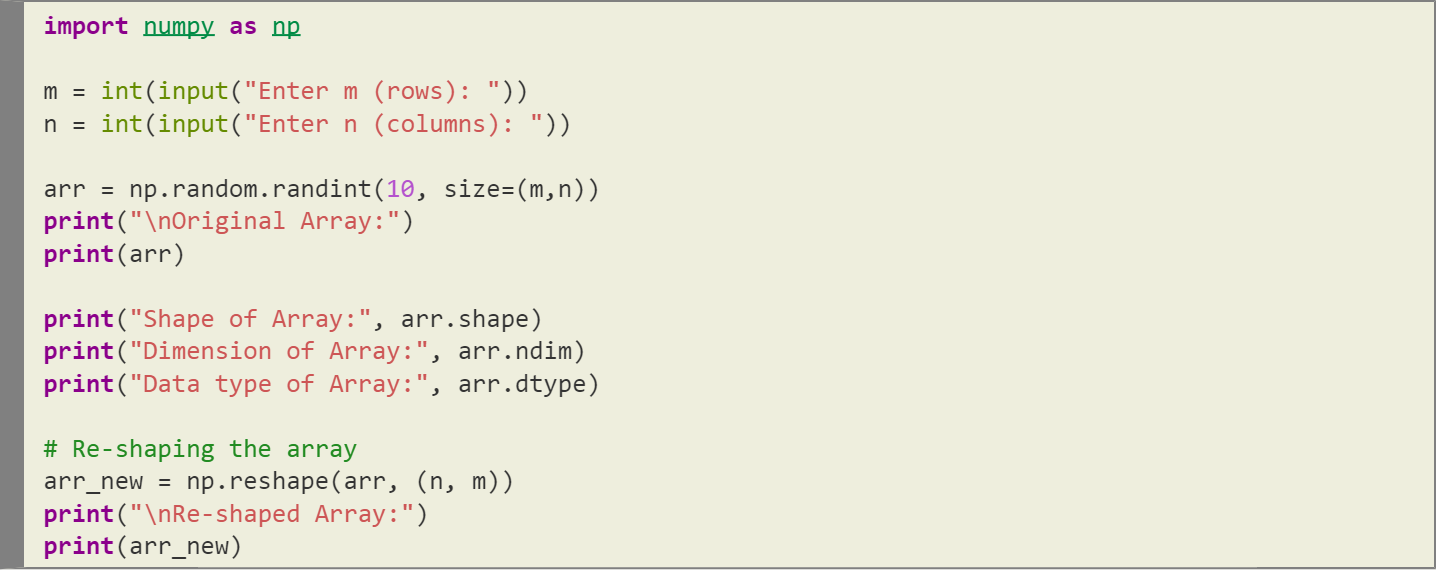
**Code:**

**Output:**

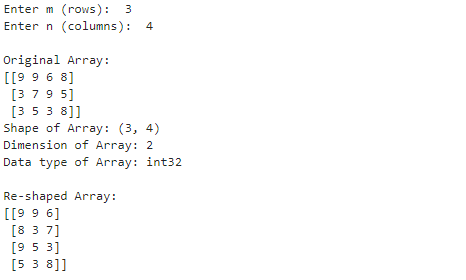


* 1. Create a 2-dimensional array of size m x n integer elements, also print the shape, type and data type of the array and then reshape it into an n x m array, where n and m are user inputs given at the run time.

**Code:**

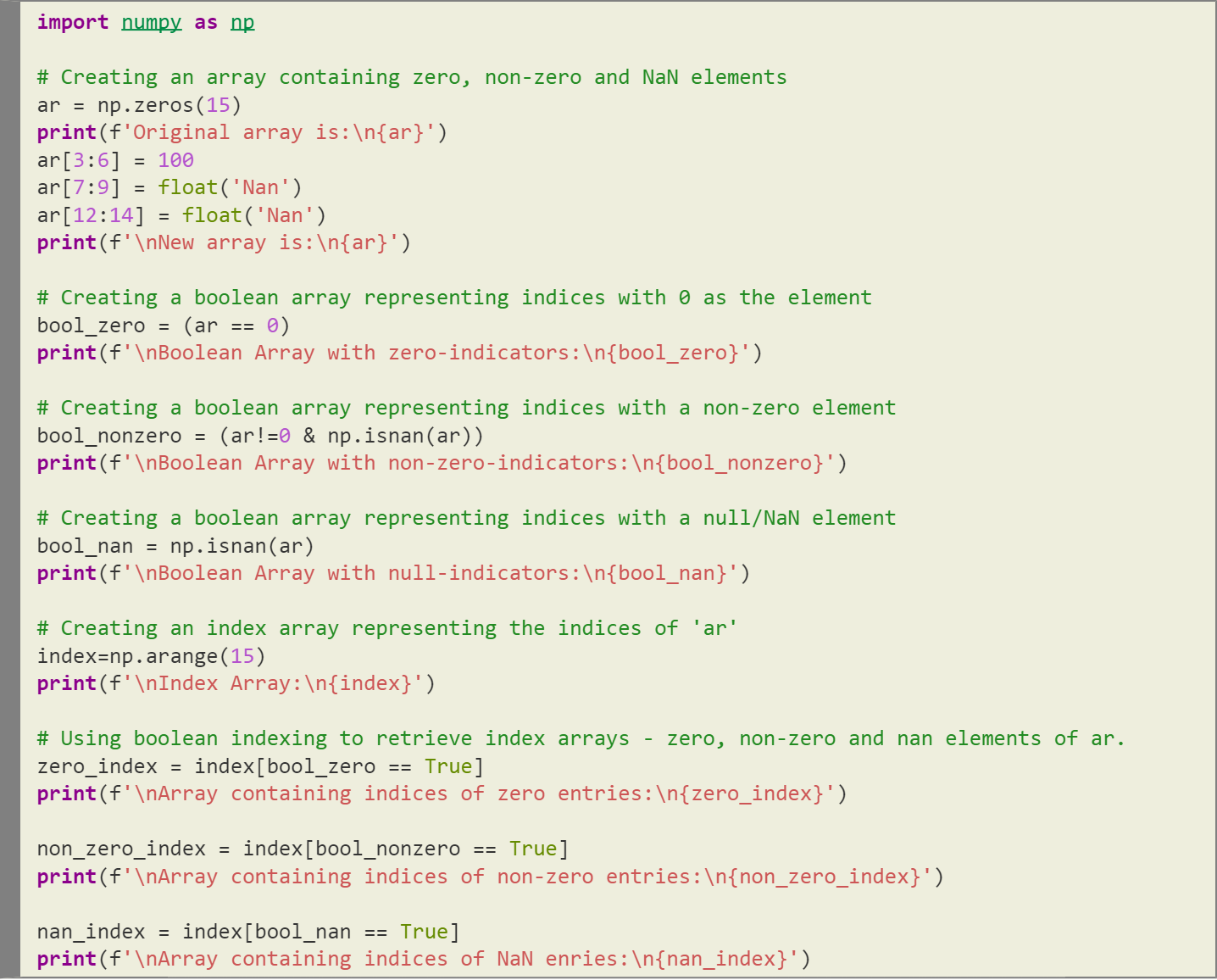


**Output:**



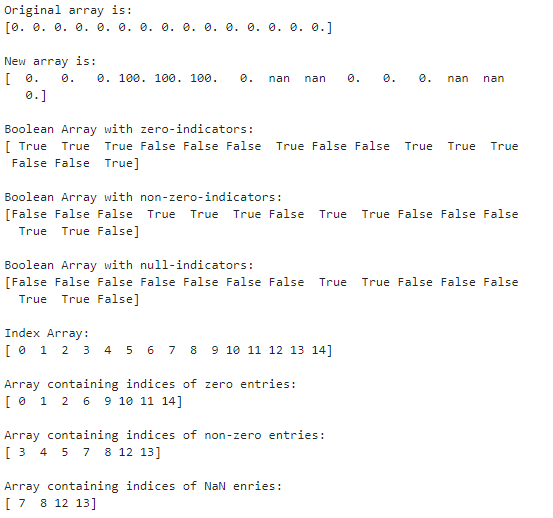
* 1. Test whether the elements of a given 1D array are zero, non-zero and NaN. Record the indices of these elements in three separate arrays.

**Code:**



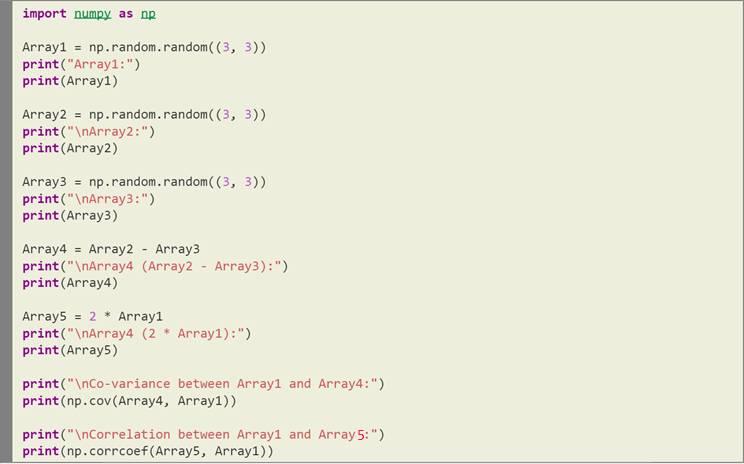
(Output on Next Page)

**Output:**



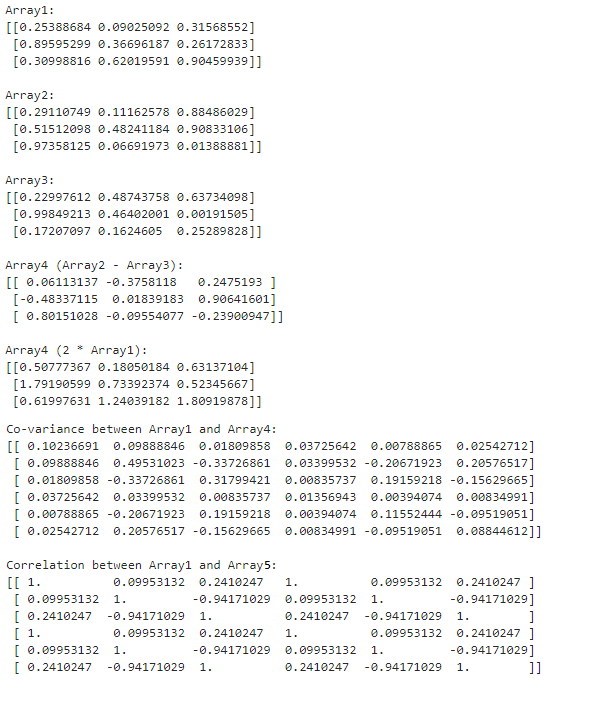
* 1. Create three random arrays of the same size: Array1, Array2 and Array3. Subtract Array 2 from Array3 and store in Array4. Create another array Array5 having two times the values in Array1. Find Co-variance and Correlation of Array1 with Array4 and Array5 respectively.

**Code:**



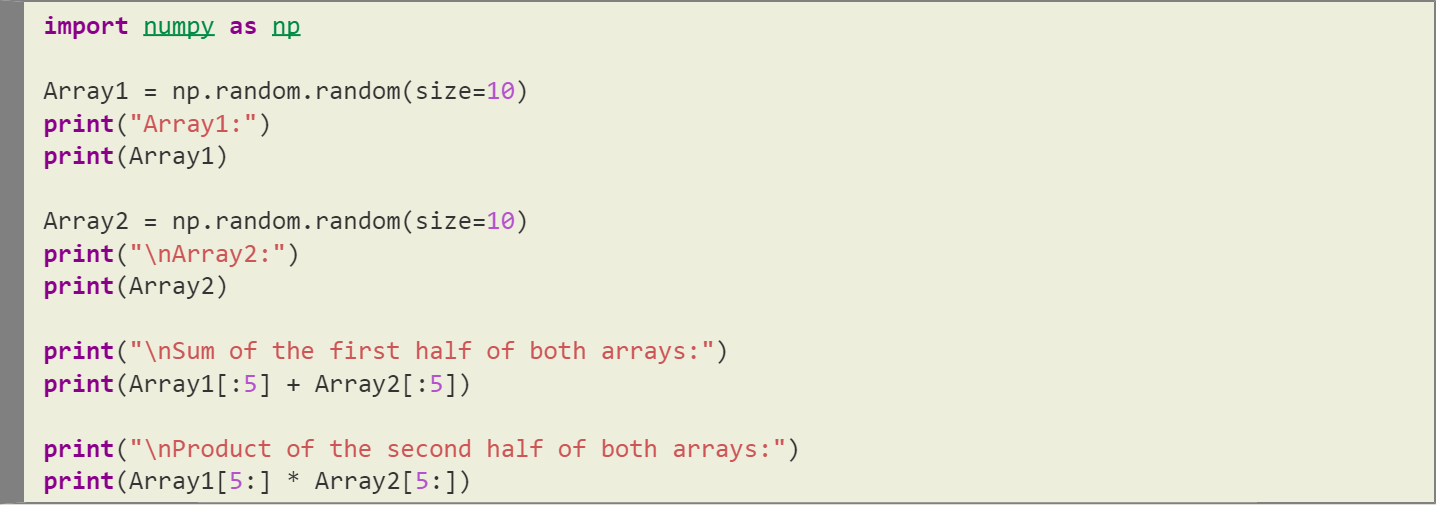
(Output on Next Page)

**Output:**

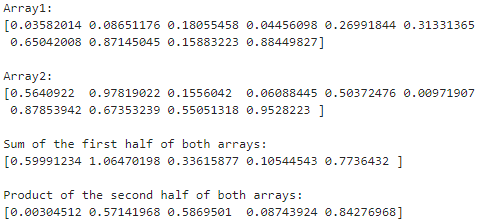


* 1. Create two random arrays of the same size 10: Array1, and Array2. Find the sum of the first half of both the arrays and product of the second half of both the arrays.

**Code:**

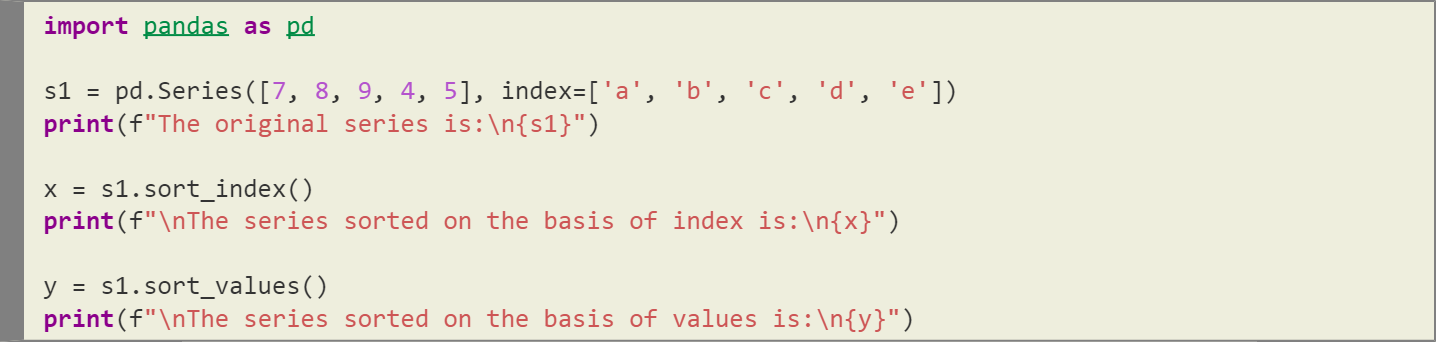


**Output:**

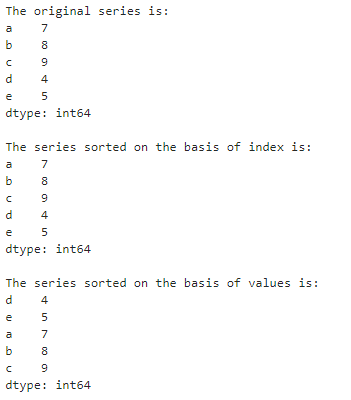


1. Do the following using PANDAS Series:
   1. Create a series with 5 elements. Display the series sorted on index and also sorted on values separately

**Code:**

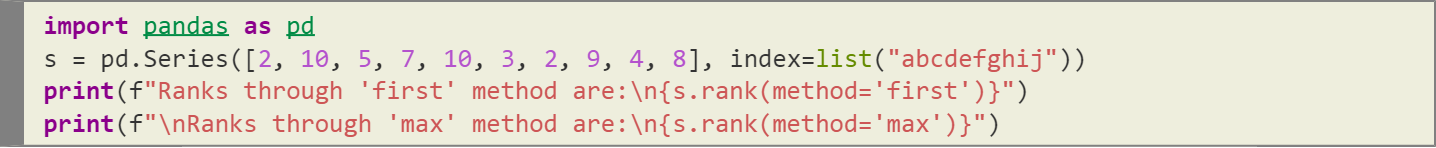


**Output:**

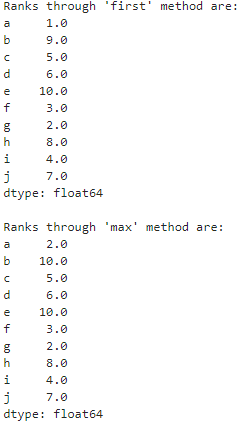


* 1. Create a series with N elements with some duplicate values. Find the minimum and maximum ranks assigned to the values using 'first' and 'max' methods

**Code:**

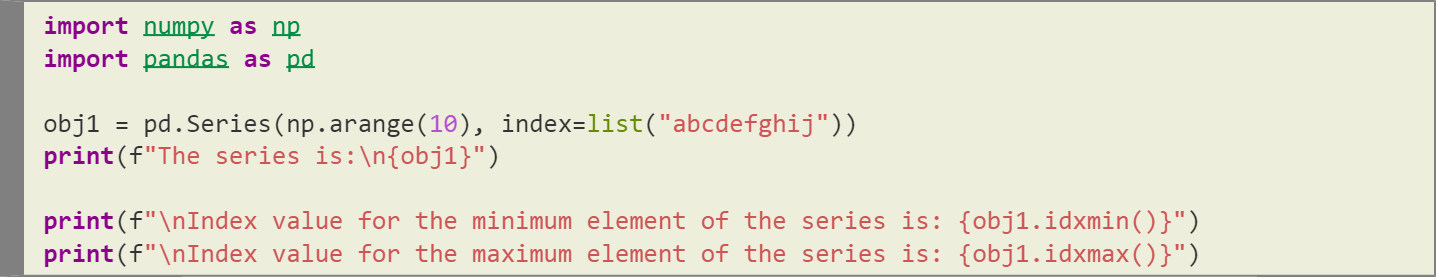


**Output:**

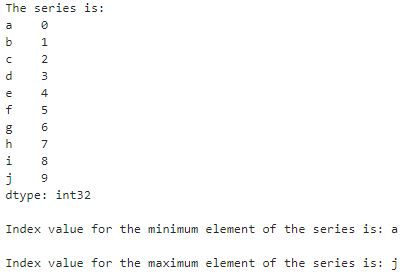


* 1. Display the index value of the minimum and maximum element of a Series

**Code:**

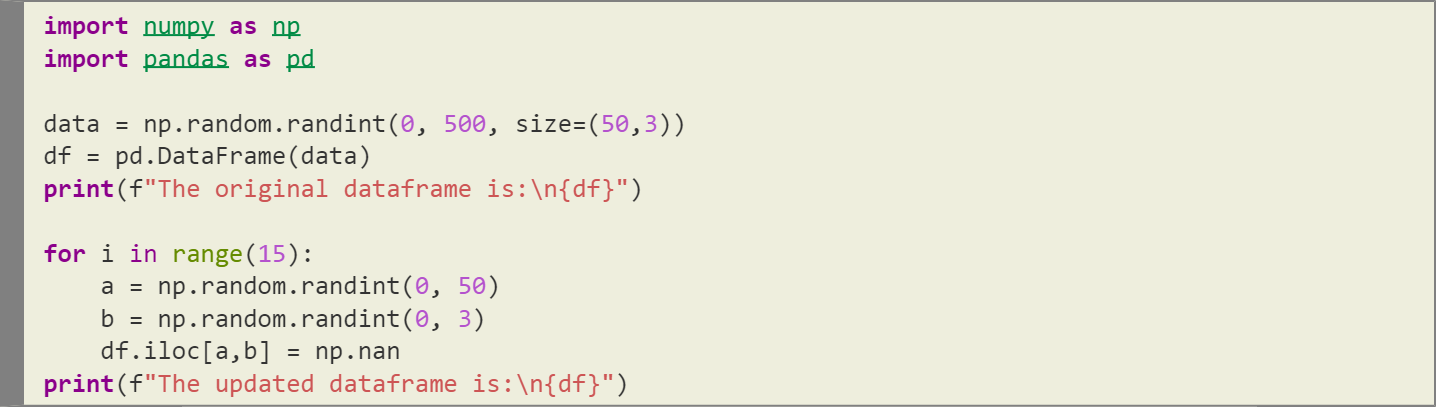


**Output:**



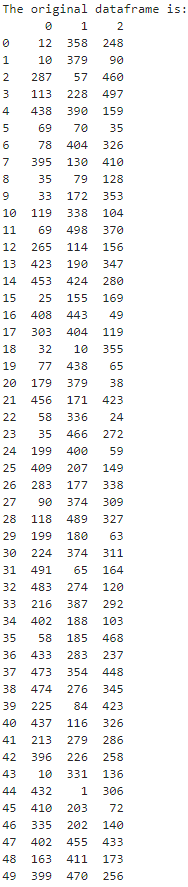
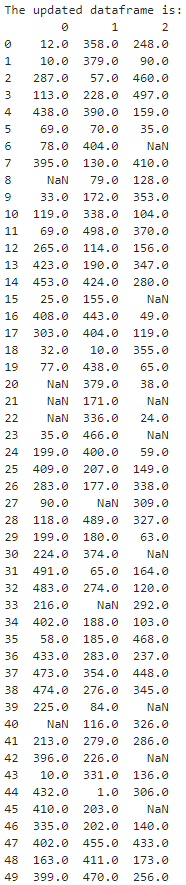
1. Create a data frame having at least 3 columns and 50 rows to store numeric data generated using a random function. Replace 10% of the values by null values whose index positions are generated using random function.

**Code:** (Creating the required data frame)



(Output on Next Page)

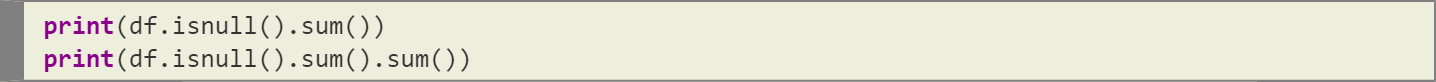
**Output:**

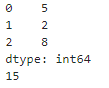
Do the following:

* 1. Identify and count missing values in a data frame.

**Code:**



**Output:**



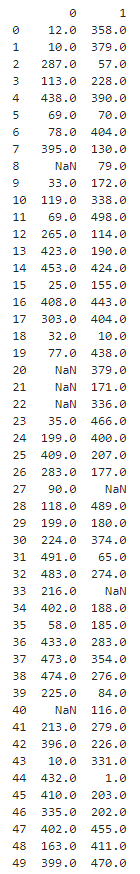
* 1. Drop the column having more than 5 null values.

**Code:**



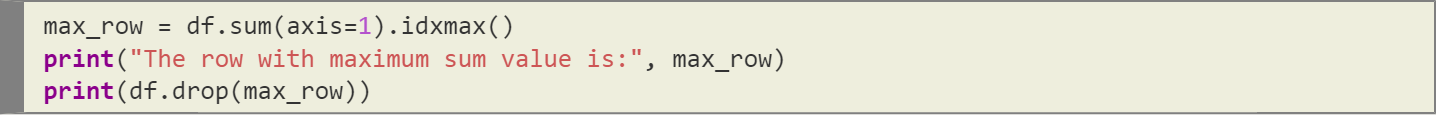
(Output on Next Page)

**Output:**



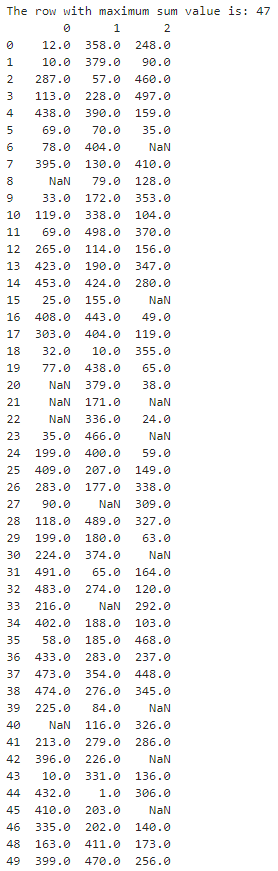
* 1. Identify the row label having maximum of the sum of all values in a row and drop that row.

**Code:**



(Output on Next Page)

**Output:**



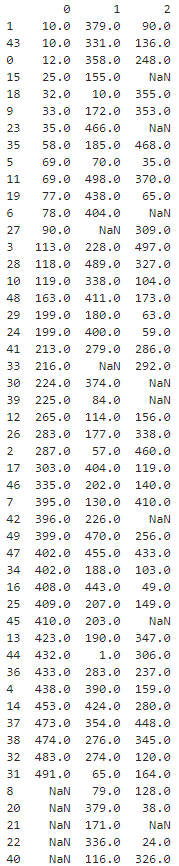
* 1. Sort the data frame on the basis of the first column.

**Code:**



(Output on Next Page)

**Output:**

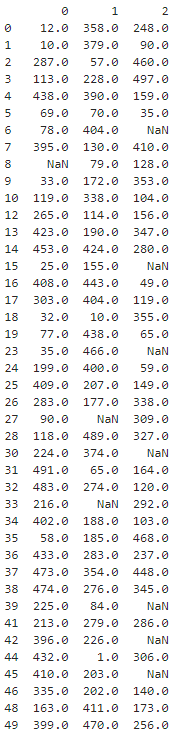


* 1. Remove all duplicates from the first column.

**Code:**

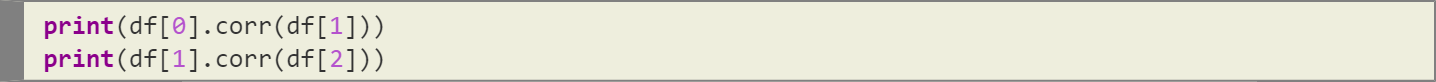


**Output:**



* 1. Find the correlation between first and second column and covariance between second and third column.

**Code:**

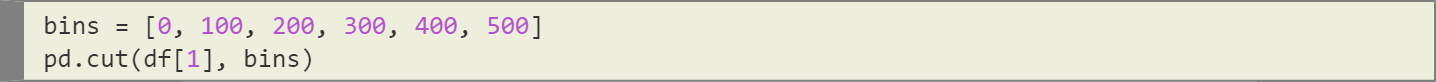


**Output:**



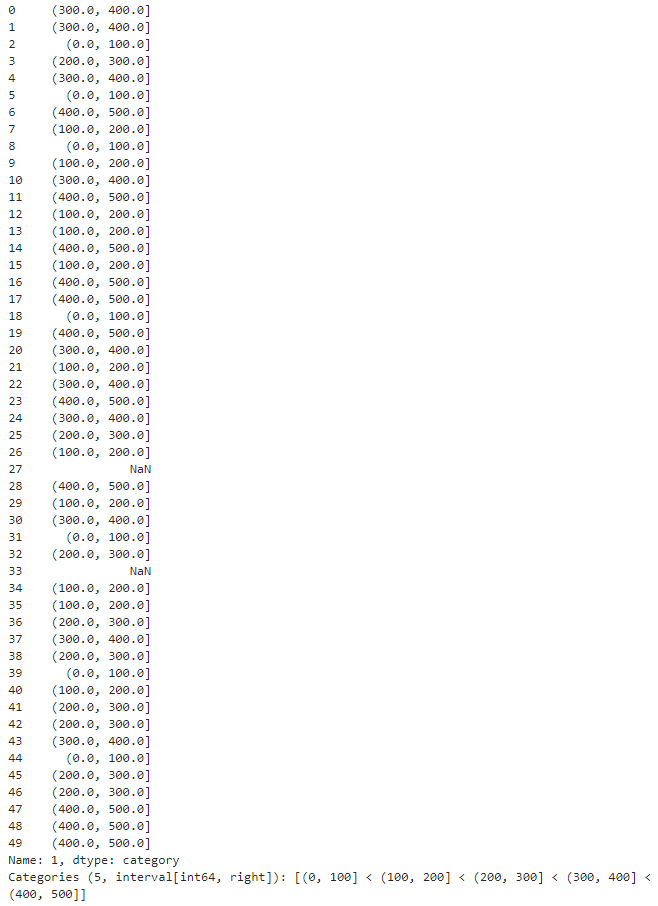
* 1. Discretize the second column and create 5 bins.

**Code:**



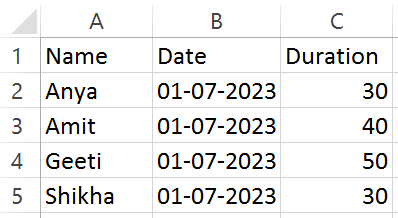
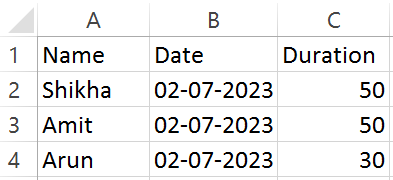
(Output on Next Page)

**Output:**

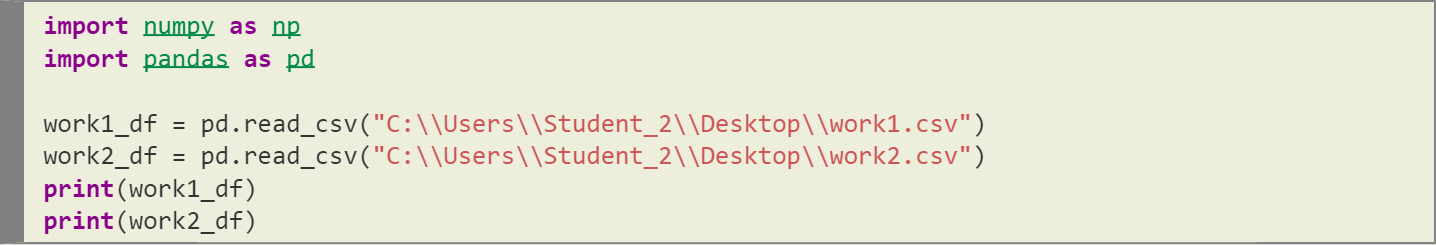


1. Consider two excel files having attendance of two workshops. Each file has three fields 'Name', 'Date', 'Duration (in minutes)' where names are unique within a file. Note that duration may take one of three values (30, 40, 50) only. Import the data into two data frames.

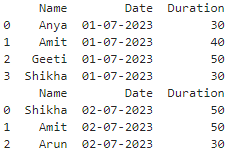
**File contents:** (**work1.csv** and **work2.csv** respectively)

**Code:** (Importing the data into data frames)



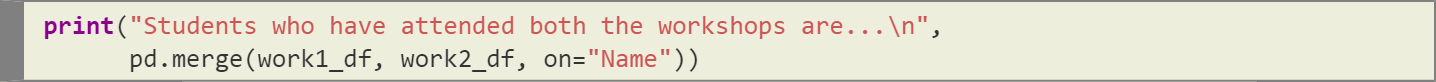
**Output:**



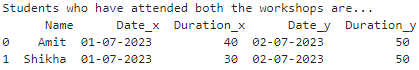
Now, do the following:

* 1. Perform merging of the two data frames to find the names of students who had attended both workshops.

**Code:**



**Output:**

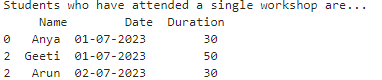


* 1. Find names of all students who have attended a single workshop only.

**Code:**

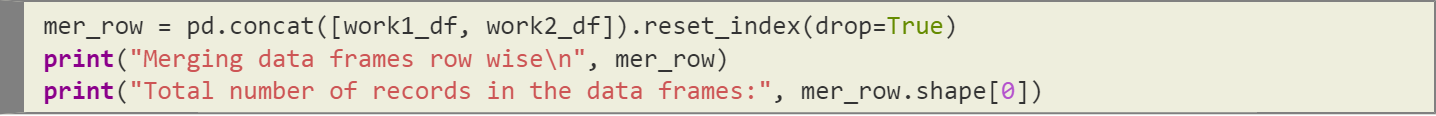


**Output:**

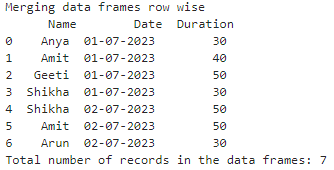


* 1. Merge two data frames row-wise and find the total number of records in the data frames.

**Code:**

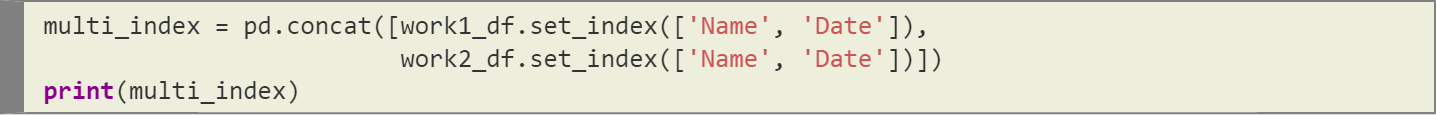


**Output:**

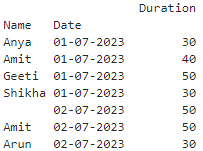


* 1. Merge two data frames row-wise and use two columns viz. names and dates as multi-row indexes.

**Code:**

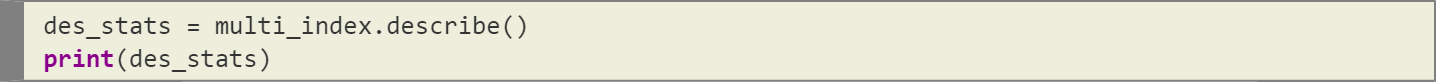


**Output:**

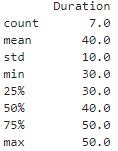


Generate descriptive statistics for this hierarchical data frame.

**Code:**



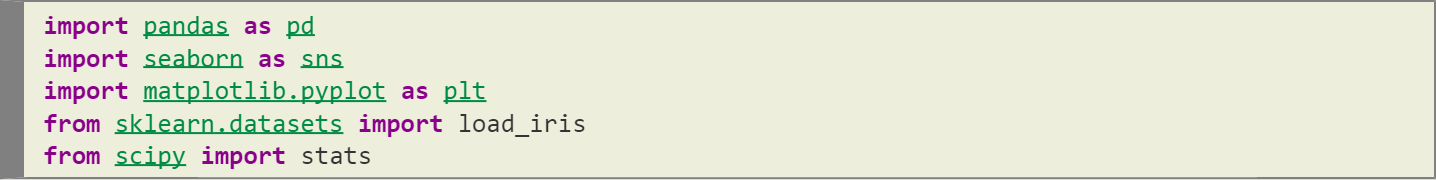
**Output:**



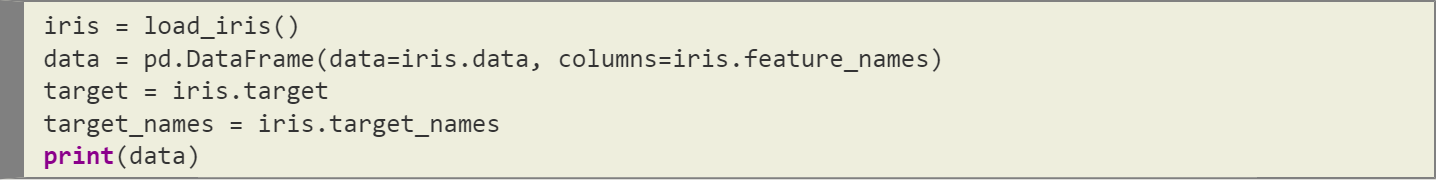
1. Using Iris data, plot the following with proper legend and axis labels: (Download IRIS data from: <https://archive.ics.uci.edu/ml/datasets/iris> or import it from sklearn datasets)

**Code:**

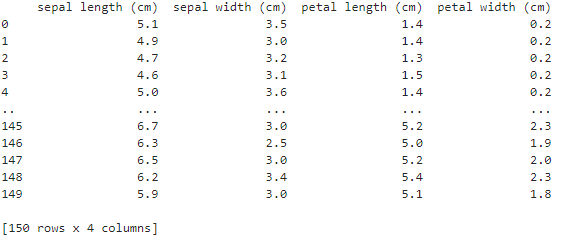
(Importing necessary libraries)



(Loading IRIS data)

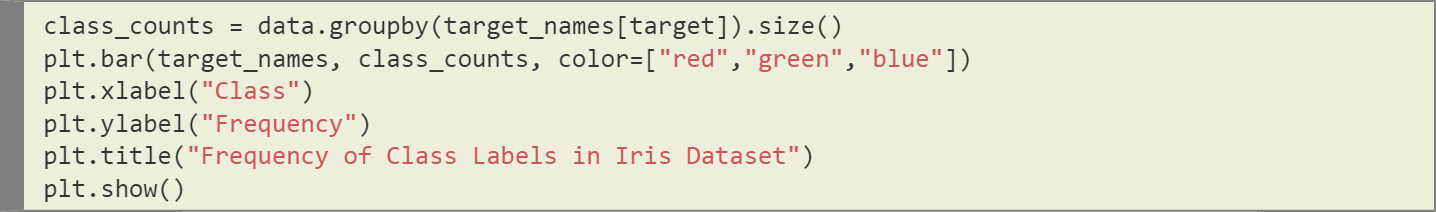


**Output:**

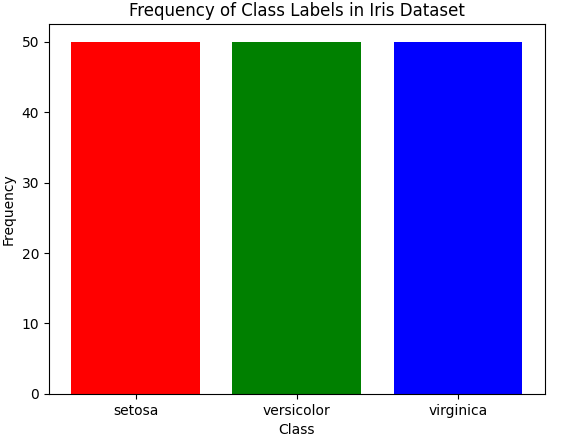


* 1. Plot bar chart to show the frequency of each class label in the data.

**Code:**

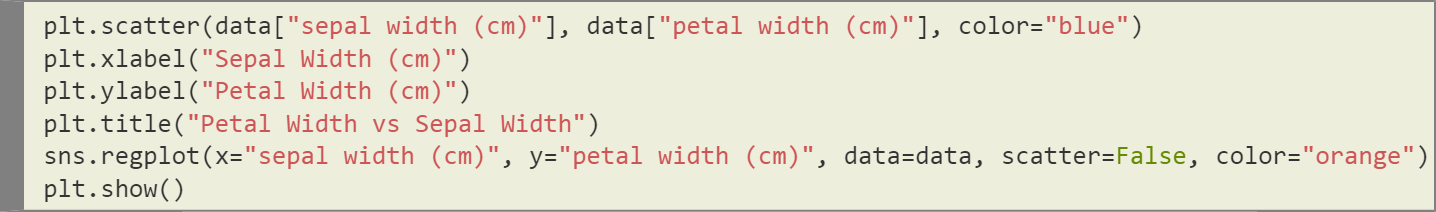


**Output:**

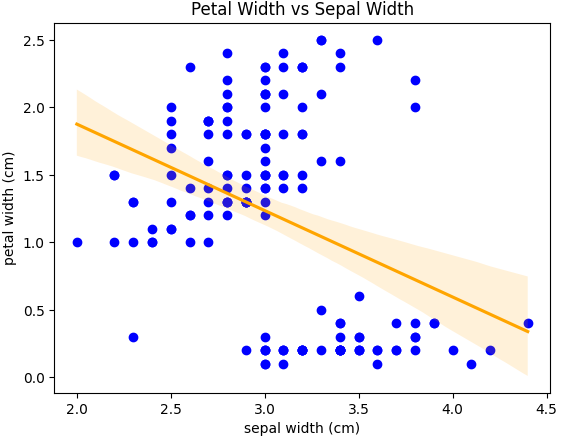


* 1. Draw a scatter plot for Petal width vs sepal width and fit a regression line

**Code:**

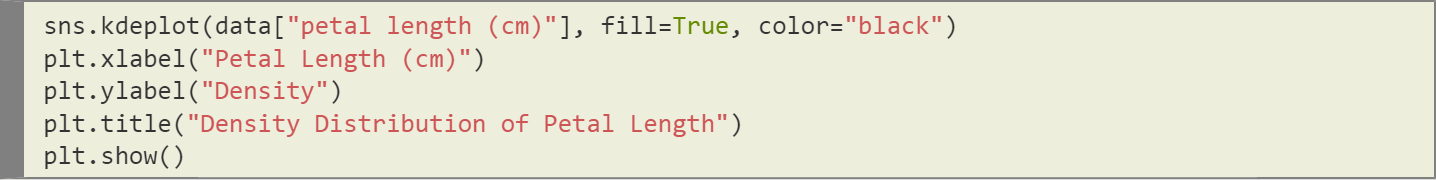


**Output:**

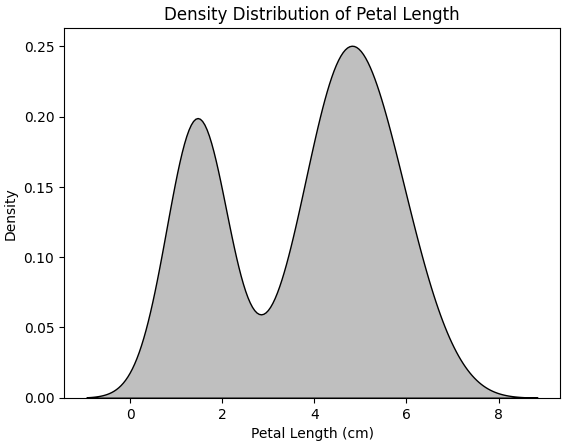


* 1. Plot density distribution for feature petal length.

**Code:**

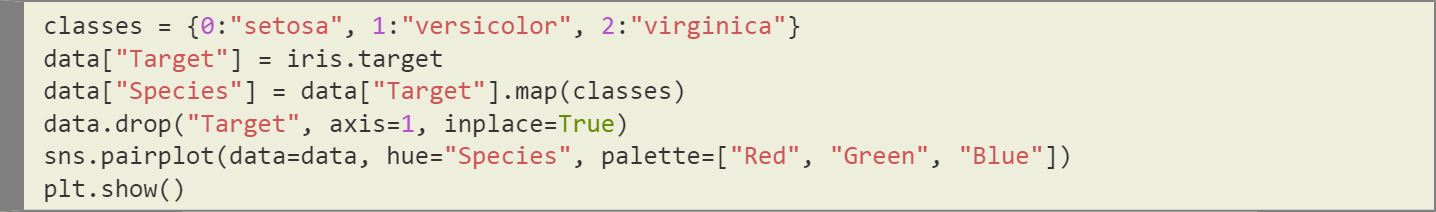


**Output:**

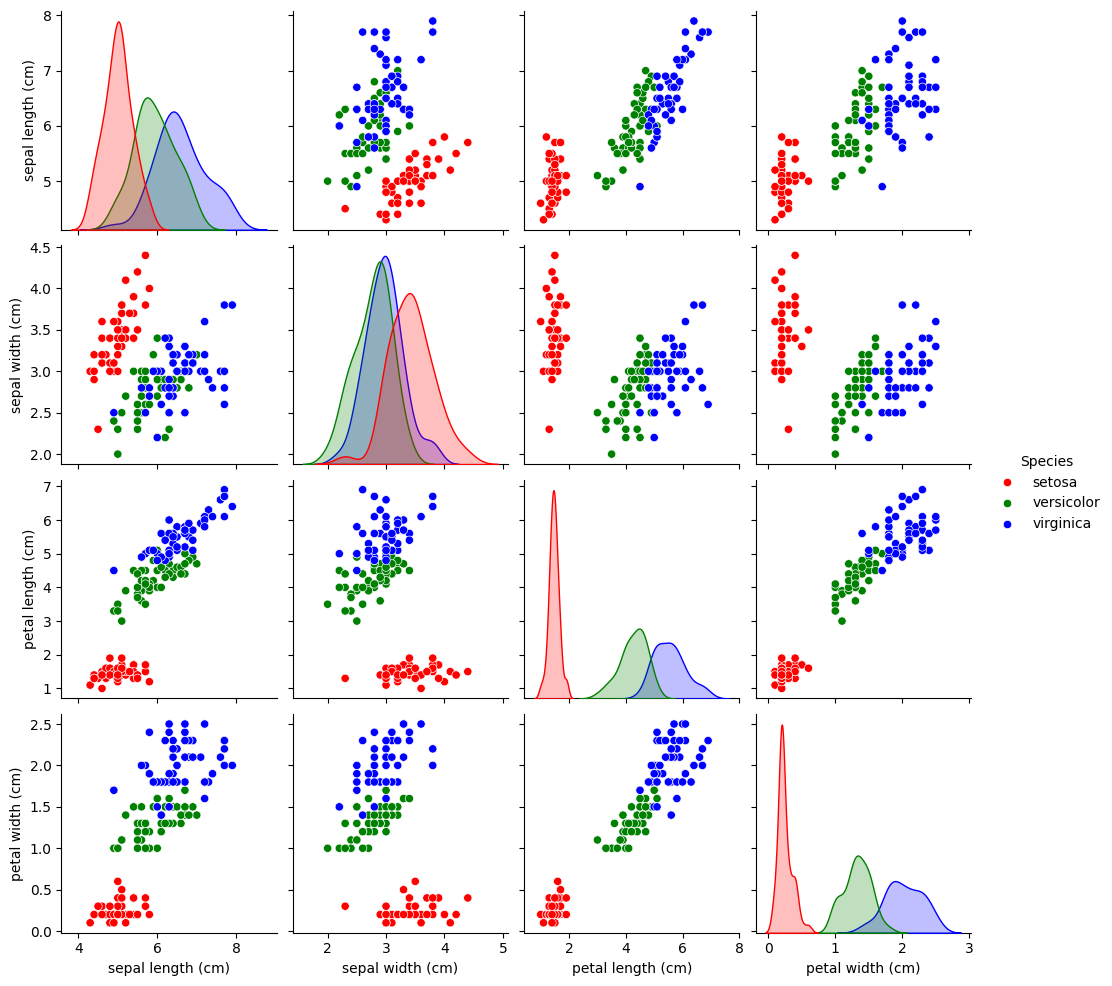


* 1. Use a pair plot to show pairwise bivariate distribution in the Iris Dataset.

**Code:**



**Output:**

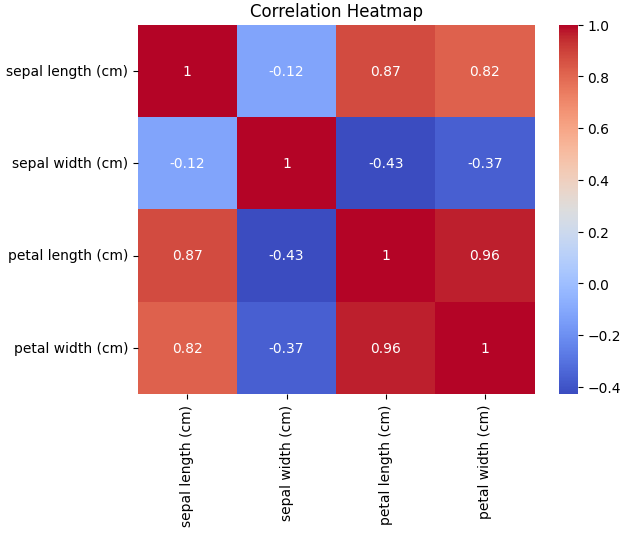


* 1. Draw heatmap for the four numeric attributes

**Code:**

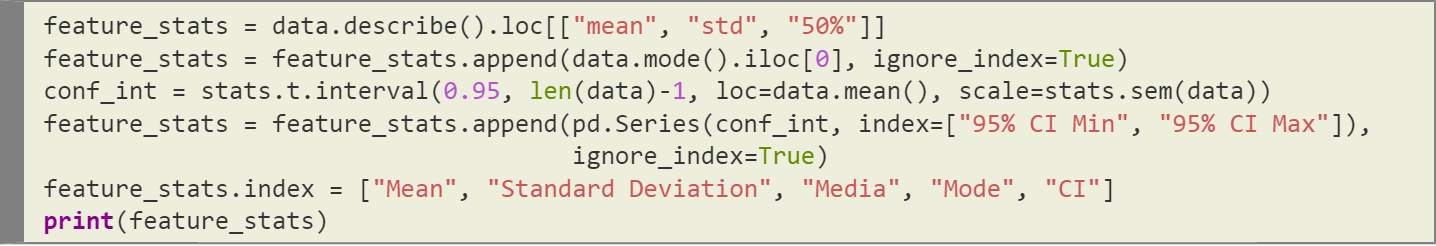


**Output:**

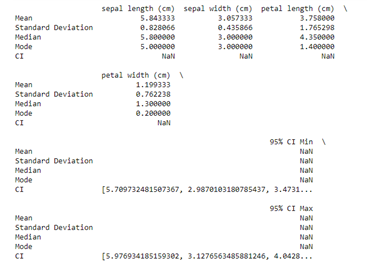


* 1. Compute mean, mode, median, standard deviation, confidence interval and standard error for each feature

**Code:**



**Output:**

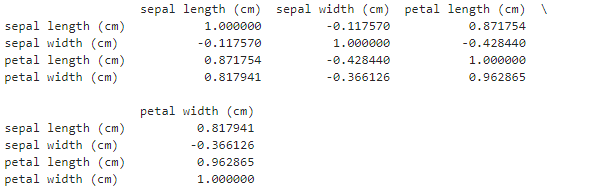


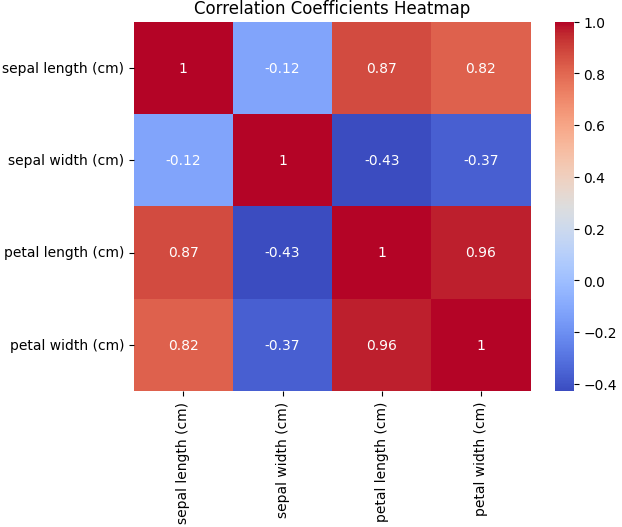
* 1. Compute correlation coefficients between each pair of features and plot heatmap

**Code:**

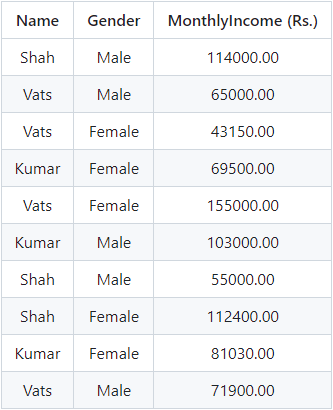


**Output:**

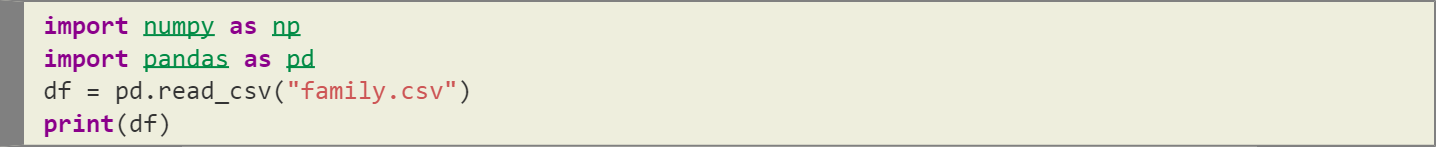




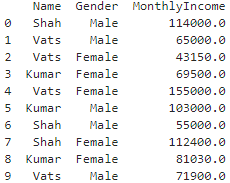
1. Consider the following data frame containing a family name, gender of the family member and her/his monthly income in each record.



**Code:** (Creating the required data frame)



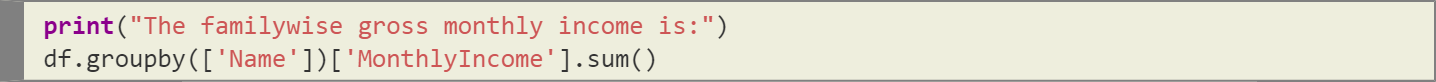
**Output:**



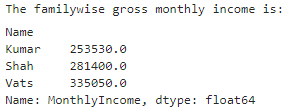
Write a program in Python using Pandas to perform the following:

* 1. Calculate and display familywise gross monthly income.

**Code:**



**Output:**

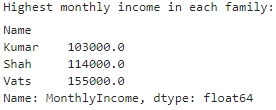


* 1. Calculate and display the member with the highest monthly income.

**Code:**



**Output:**

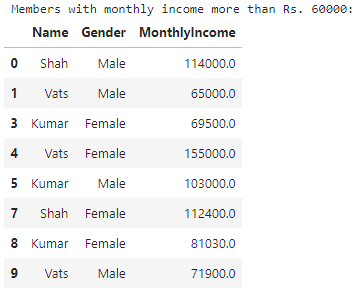


* 1. Calculate and display monthly income of all members with income greater than Rs. 60000.00.

**Code:**

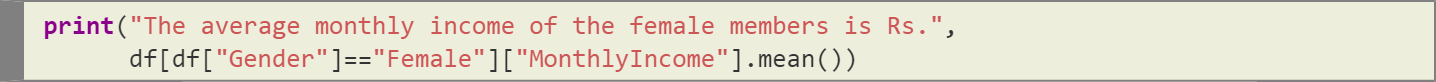


**Output:**



* 1. Calculate and display the average monthly income of the female members

**Code:**

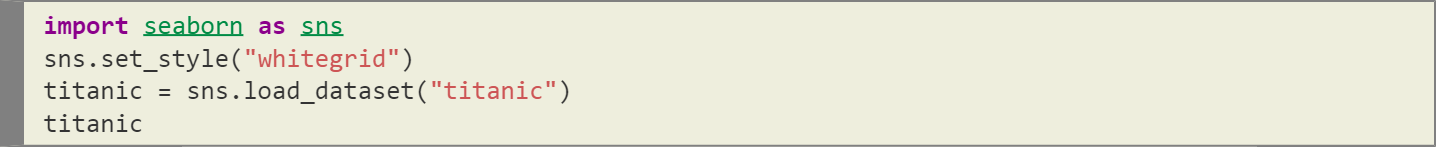


**Output:**

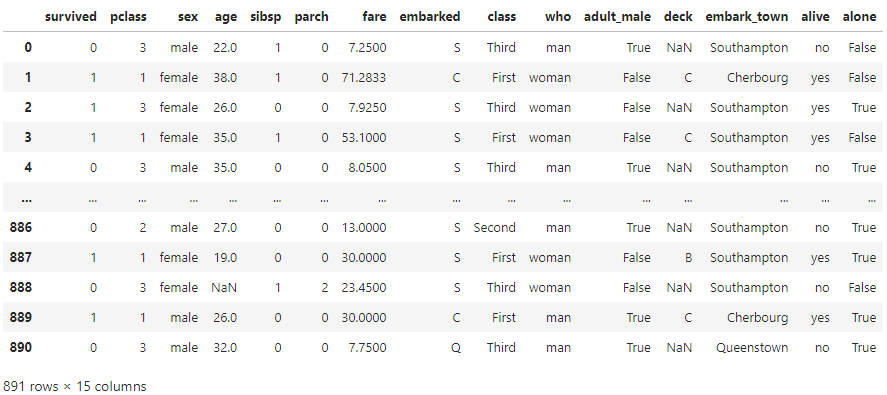


1. Using Titanic dataset, to do the following:

**Code:** (Importing **seaborn** and loading TITANIC data)



**Output:**



* 1. Find total number of passengers with age less than 30

**Code:**



**Output:**



* 1. Find total fare paid by passengers of first class

**Code:**

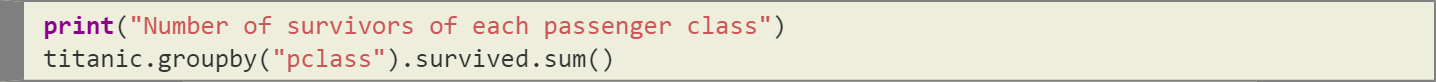


**Output:**

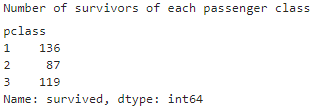


* 1. Compare number of survivors of each passenger class

**Code:**



**Output:**



* 1. Compute descriptive statistics for any numeric attribute genderwise

**Code:**



**Output:**

