Teaching Plan

Course: ALL STREAMS

Semester-II

Subject: GE2a: Data Analysis and Visualization using Python

Teaching Plan	
Weeks	**Topics**
Week 1 to 3	Unit 1 Introduction to basic statistics and analysis: Fundamentals of Data Analysis, Statistical foundations for Data Analysis, Types of data, Descriptive Statistics, Correlation and covariance, Linear Regression, Statistical Hypothesis Generation and Testing Python Libraries: NumPy, Pandas, Matplotlib
Week 4 to 6	Unit 2 Array manipulation using Numpy: NumPy array: Creating NumPy arrays, various data types of NumPy arrays Indexing and slicing, swapping axes, transposing arrays, data processing using Numpy arrays
Week 7 to 9	Unit 3 Data Manipulation using Pandas: Data Structures in Pandas: Series, Data Frame, Index objects, loading data into Panda's data frame, Working with Data Frames: Arithmetics, Statistics, Binning, Indexing.
Week 10 to 12	Reindexing, Filtering, Handling missing data, Hierarchical indexing, Data wrangling: Data cleaning, transforming, merging and reshaping.
Week 13 to 15	Unit 4 Plotting and Visualization: Using Matplotlib to plot data: figures, subplots, markings, color and line styles, labels and legends, Plotting functions in Pandas: Lines, bar, Scatter plots, histograms, stacked bars, Heatmap.

References

1. McKinney W. *Python for Data Analysis: Data Wrangling with Pandas*, *NumPy and IPython*. 2nd edition. O'Reilly Media, 2018..

2. Molin S. *Hands-On Data Analysis with Pandas*, Packt Publishing, 2019.

3. Gupta S.C., Kapoor V.K., *Fundamentals of Mathematical Statistics*, Sultan Chand & Sons, 2020.

Practical List For Data Analysis and Visualization GE Sem II

Note:

● Any platform for Python can be used for lab exercises

● Use a data set of your choice from Open Data Portal (https:// data.gov.in/, UCI repository) or load from scikit, seaborn library for the following exercises to practice the concepts learnt.

1. Write programs in Python using NumPy library to do the following:

a. Compute the mean, standard deviation, and variance of a two dimensional random integer array along the second axis.

b. 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.

c. 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.

d. 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.

e. 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.

2. Do the following using PANDAS Series:

a. Create a series with 5 elements. Display the series sorted on index and also sorted on values separately

b. 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

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

3. 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. Do the following:

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

b. Drop the column having more than 5 null values.

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

d. Sort the data frame on the basis of the first column.

e. Remove all duplicates from the first column.

f. Find the correlation between first and second column and covariance between second and third column.

g. Discretize the second column and create 5 bins.

4. Consider two excel files having attendance of two workshos. 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 and do the following:

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

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

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

d. Merge two data frames row-wise and use two columns viz. names and dates as multi-row indexes. Generate descriptive statistics for this hierarchical data frame.

5. 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)

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

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

c. Plot density distribution for feature petal length.

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

e. Draw heatmap for the four numeric attributes

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

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

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

Name	Gender	MonthlyIncome (Rs.)
Shah	Male	114000.00
Vats	Male	65000.00
Vats	Female	43150.00
Kumar	Female	69500.00
Vats	Female	155000.00
Kumar	Male	103000.00
Shah	Male	55000.00
Shah	Female	112400.00
Kumar	Female	81030.00
Vats	Male	71900.00

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

a. Calculate and display familywise gross monthly income.

b. Calculate and display the member with the highest monthly income.

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

d. Calculate and display the average monthly income of the female members.

7. Using Titanic dataset, to do the following:

a. Find total number of passengers with age less than 30

b. Find total fare paid by passengers of first class

c. Compare number of survivors of each passenger class

d. Compute descriptive statistics for any numeric attribute genderwise.