Department of Computer Science

COURSES OFFERED BY DEPARTMENT OF COMPUTER SCIENCE

(Computer Science Courses for Undergraduate Programme of study with Computer Science discipline as one of the three Core Disciplines)

Semester 4

DISCIPLINE SPECIFIC Elective - (DSE-2a) : Data Exploration and Visualization

CREDIT DISTRIBUTION, ELIGIBILITY AND PRE-REQUISITES OF THE COURSE

Course title & Code	Credits	Credit distribution of the course			Eligibility criteria	Pre-requisite of the course (if any)
		Lecture	Tutorial	Practical/ Practice		
DSE 02a Data Exploration and Visualization	**4**	**3**	**0**	**1**	Pass in Class XII	Basic Knowledge of Python

Learning Objectives

The course is designed to equip students with the skills to analyze diverse real-world data to extract data insights and interpret results. Subsequently, they will learn to apply exploratory data science techniques to effectively communicate findings, using Python for effective problem-solving.

Learning Outcomes

On successful completion of the course, students will be able to:

- Create and manipulate NumPy arrays to perform data analysis.
- Use Pandas methods to import, export, and preprocess data from various sources.
- Perform basic data manipulation tasks, including data cleaning, filtering, sorting, and merging on Pandas objects .

- Use grouping and aggregation operations in Pandas to summarize data in Series and DataFrame objects; and analyze and interpret data based on grouped and aggregated results.
- Use Matplotlib and Seaborn to create static visualizations and plotly to create interactive visualizations of data to communicate data insights.

SYLLABUS OF DSE 02 a

Unit 1 (10 Hour): Creating and Manipulating NumPy arrays: creating arrays, indexing and slicing, mathematical operations with NumPy arrays

Unit 2(15 Hours): Data Manipulation with Pandas: Series and DataFrame objects; importing and exporting data from various file formats into pandas DataFrame; Data selection and filtering-indexing, slicing, conditional filtering using boolean indexing; Data Cleaning- handling missing data in Pandas and outlier detection; Data Manipulation-sorting, reshaping, merging.

Unit 3 (5 Hours): Grouping and Aggregation with Pandas: Grouping data using Pandas, applying aggregation functions such as sum, mean, count, etc.to grouped data, using pivot tables and cross-tabulation for data summarization

Unit 4 (10 Hours): Data Visualization with Matplotlib and Seaborn: Introduction to Matplotlib and Seaborn to plot data using figures and subplots, Plots - Line plots, scatter plots, and bar plots, Visualizing distributions using histogram and box plots, Customizing plot aesthetics and adding annotations

Unit 5 (5 Hours): Interactive Visualizations with Plotly: Introduction to Plotly library for interactive visualization; Creating interactive line plots, scatter plots, and bar plots; Adding interactivity with hover effects, zooming, and panning

Essential/recommended readings

1. VanderPlas, J. Python data science handbook: Essential tools for working with data. " O'Reilly Media, Inc.", 2nd edition.

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

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

4. Rahman K. Python Data Visualization Essentials Guide: Become a Data Visualization expert by building strong proficiency in Pandas, Matplotlib, Seaborn, Plotly, Numpy, and Bokeh, BPB 2021

Additional References:

1. Chen D. Y, Pandas for Everyone: Python Data Analysis, Pearson, 2018.

Online references/material:

1. https://www.indeepdata.com/blog/exploratory-data-analysis/

Suggested Practical List (If any): (30 Hours)

Use 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 a program using the NumPy library to perform the following tasks:

   A. Generate a 5x2 integer array with values ranging from 50 to 100, where each element has a difference of 5. Reshape the resulting array to a size of 10x1.
   B. Create a 1D random array with values ranging from 1 to 100. Calculate various statistical measures such as minimum, maximum, mean, median, standard deviation, number of unique values, count of unique values, and the most frequent value in the array.
   C. Create a 5x5 identity matrix where all the diagonal elements are set to the value 5.
   D. Consider a dataset containing the heights (in centimeters) and weights (in kilograms) of 20 individuals. Your task is to perform various operations using the NumPy library to analyze the data.
      a. Create a NumPy array called "heights" with the following height values: [165, 170, 175, 168, 172, 180, 160, 169, 176, 171, 174, 182, 158, 167, 173, 179, 163, 166, 177, 181]. Create a NumPy array called "weights" with the following weight values: [60, 65, 70, 75, 80, 85, 55, 58, 63, 68, 72, 77, 50, 62, 67, 74, 52, 57, 69, 73].
      b. Create a new NumPy array called "combined" by stacking the heights and weights arrays such that the shape of the resulting array is 20 x 2.
      c. Calculate and print the mean height and weight of the individuals in the dataset.
      d. Find and print the index of the shortest and tallest individuals in the dataset.
      e. Sort the array based on height on the individuals.
      f. Swap the positions of the two columns in the array.
      g. Retrieve records of individuals having weight below 70kg.

2. Write a program using the Pandas library to perform the following operations on the penguins dataset from the Seaborn library:

A. Load the penguins dataset into a Pandas dataframe.
B. Determine the number of observations/records and the number of attributes in the dataframe.
C. Display the names of the attributes, row indexes, and data types of each attribute in the dataframe.
D. Display the first 5 and last 5 records of the dataframe.
E. Retrieve the values of the second column for the third and fourth records.
F. Display a summary of the data distribution for all attributes in the dataframe.
G. Compute the pairwise correlation between all attributes in the dataframe.

3. Consider the Titanic dataset, which contains information about passengers on board the Titanic, including their age, gender, passenger class, survival status, and other attributes. Write a program using the Pandas library to perform the following operations on the Titanic dataset:

A. Load the Titanic dataset into a Pandas DataFrame.
B. Check for any duplicate records and missing values in the dataset and handle them appropriately.
C. Calculate and display the total number of passengers who survived and those who did not.
D. Filter the DataFrame to select only the records of passengers who were under the age of 18.
E. Calculate the average age for passengers belonging to each of the passenger class.
F. Create a new column in the DataFrame called "Family Size" that represents the total number of family members (including the passenger) on board.
G. Calculate the correlation between age and fare attributes of the dataset.
H. Create a contingency table that shows the count of passengers based on their survival status (survived or not) and passenger class (first, second, or third class). for titanic dataset

4. Utilize the iris dataset from the Sklearn library to generate various visual representations of the data using the Matplotlib and or Seaborn libraries with proper legends and labels. Perform the following tasks:

A. Create a scatter plot to visualize the relationship between petal length and petal width for different instances of iris flowers.
B. Generate histograms to display the data distribution of each of the four attributes in the iris dataset.
C. Construct a pie chart to illustrate the frequency count of each flower type in the iris dataset.

D. Create a pair plot that showcases the relationship between every pair of attributes in the iris dataset (only seaborn library).

5. Create the visualizations of question 4 using plotly library.

DISCIPLINE SPECIFIC Elective - (DSE-2b) : Software Engineering

Semester 4

CREDIT DISTRIBUTION, ELIGIBILITY AND PRE-REQUISITES OF THE COURSE

Course title & Code	Credits	Credit distribution of the course			Eligibility criteria	Pre-requisite of the course (if any)
		Lecture	Tutorial	Practical/ Practice		
DSE 02b Software Engineering	4	3	0	1	Pass in Class XII	**Knowledge of any programming Language**

Learning Objectives

The course introduces software development life cycle and software project management. It includes requirement engineering, software project planning, software design, software quality and software testing.

Learning Outcomes

On successful completion of the course, students will be able to:

- Understand and apply the software process models.
- Elicitate and analyse customer requirements and map requirements to design models.
- Estimate size, effort and cost required to build software.
- Identify risks involved in the software development and understand software quality.
- Design test cases to perform software testing.