

Suggestive readings

- (i) Sahni, S., *Data Structures, Algorithms and applications in C++*, 2nd edition, Universities Press, 2011.
- (ii) Langsam Y., Augenstein, M. J., & Tanenbaum, A. M. *Data Structures Using C and C++*, Pearson, 2009.

Note: Examination scheme and mode shall be as prescribed by the Examination Branch, University of Delhi, from time to time.

DISCIPLINE SPECIFIC CORE COURSE – A2 : DATA INTERPRETATION AND VISUALIZATION USING PYTHON

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		
A2: Data Interpretation and Visualization using Python	4	3	0	1	Class XII pass with Mathematics	knowledge of Python

Learning Objectives

This course is designed to introduce the students to the application of Python to get a deterministic view of data and interpret results..

Learning outcomes

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

- Interpret Data
- Obtain a deterministic view of data
- Perform data handling using Numpy arrays
- Load, clean, transform, merge and reshape data using Pandas
- Visualize data using Pandas and matplotlib libraries

SYLLABUS OF A2

UNIT – I (06 Hours)

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

UNIT – II (09 Hours)

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

UNIT – III (12 Hours)

Data Manipulation using Pandas: Data Structures in Pandas: Series, DataFrame, Index objects, Loading data into Pandas data frame, Working with Data Frames: Arithmetics, Statistics, Binning, Indexing, Reindexing, Filtering, Handling missing data, Hierarchical indexing, Data wrangling: Data cleaning, transforming, merging and reshaping

UNIT – IV (12 Hours)

Plotting and Visualization: Using Matplotlib to plot data: figures, subplots, markings, color and line styles, labels and legends, plotting functions in Pandas: Line, bar, Scatter plots, histograms, stacked bars, Heatmap

UNIT-V (06 Hours)

Data Aggregation and Group operations: Group by Mechanics, Data aggregation, General split-apply-combine, Pivot tables and cross tabulation.

Practical component (if any) – 30 Hours

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

1. Load a Pandas dataframe with a selected dataset. Identify and count the missing values in a dataframe. Clean the data after removing noise as follows
 - a. Drop duplicate rows.
 - b. Detect the outliers and remove the rows having outliers
 - c. Identify the most correlated positively correlated attributes and negatively correlated attributes
2. Import iris data using sklearn library or (Download IRIS data from: <https://archive.ics.uci.edu/ml/datasets/iris> or import it from sklearn.datasets)

- i. Compute mean, mode, median, standard deviation, confidence interval and standard error for each feature
 - ii. Compute correlation coefficients between each pair of features and plot heatmap
 - iii. Find covariance between length of sepal and petal
 - iv. Build contingency table for class feature
3. Load Titanic data from sklearn library, plot the following with proper legend and axis labels:
 - a. Plot bar chart to show the frequency of survivors and non-survivors for male and female passengers separately
 - b. Draw a scatter plot for any two selected features
 - c. Compare density distribution for features age and passenger fare
 - d. Use a pair plot to show pairwise bivariate distribution
4. Using Titanic dataset, 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
5. Download any dataset and do the following
 - a. Count number of categorical and numeric features
 - b. Remove one correlated attribute (if any)
 - c. Display five-number summary of each attribute and show it visually

Essential/recommended readings

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*, 12th edition, Sultan Chand & Sons, 2020.

Suggestive readings

- (i) Chen D. Y. *Pandas for Everyone: Python Data Analysis*, 1st edition, Pearson Education, 2018.
- (ii) Miller J.D. *Statistics for Data Science*, Packt Publishing Limited, 2017.

Note: Examination scheme and mode shall be as prescribed by the Examination Branch, University of Delhi, from time to time.