DSC-A2/GE2a/DSE: DATA ANALYSIS 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
		Lecture	Tutorial	Practical/ Practice		
Data Analysis and Visualization using Python	**4**	**3**	**0**	**1**	**Class XII pass with Mathematics**	Programming using Python

Course Objectives

This course is designed to introduce the students to real-world data analysis problems, the use of statistics to get a deterministic view of data, and interpreting results in the field of exploratory data science using Python. This course is the first in the “Data Science” pathway and builds the foundation for three subsequent courses in the pathway.

Learning outcomes

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

1. Apply descriptive statistics to obtain a deterministic view of data

2. Perform data handling using Numpy arrays

3. Load, clean, transform, merge, and reshape data using Pandas

4. Visualize data using Pandas and matplot libraries

5. Solve real world data analysis problems

SYLLABUS OF DSE

Unit 1 (**10 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, Seaborn

Unit 2 **(8 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 3 **(12 hours)**

Data Manipulation using Pandas:

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

Unit 4 **(8 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, 3D Plotting, interactive plotting using Bokeh and Plotly.

Unit 5 **(7 hours)**

Data Aggregation and Group operations:

Group by mechanics, Data aggregation, General split-apply-combine, Pivot tables and cross tabulation

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

Additional References

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

2. Miller J.D. Statistics for Data Science, Packt Publishing Limited, 2017.

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

Practical exercises such as

Use a dataset 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. 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

Project: Students are encouraged to work on a good dataset in consultation with their faculty and apply the concepts learned in the course.

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

DSC11/DSC05/GE3a: DATABASE MANAGEMENT SYSTEMS

Credit distribution, Eligibility and Prerequisites 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		
Database Management Systems	**4**	**3**	**0**	**1**	Pass in Class XII	NIL

Course Objectives

The course introduces the students to the fundamentals of database management system and its architecture. Emphasis is given on the popular relational database system including data