DSE – 13
Data Analysis 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		
Data Analysis and Visualization	**4**	**3**	**0**	**1**	**Class XII pass with Mathematics**	DSC-01(Programming using Python), Probability and Statistics, Linear Algebra

Learning Objectives:

1. *To introduce the students to real-world data analysis problems.*
2. *Students will be able to use statistics to get a deterministic view of data and interpret results in the field of exploratory data science using Python.*

Learning Outcome:

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 usingPandas
4. Visualize data using Pandas and matplotlibraries

UNIT-I **(5 hours)**

Introduction to basic statistics and analysis: Fundamentals of Data Analysis, Statistical foundations for Data Analysis, Typesof data, Descriptive Statistics, Correlation and covariance, Linear Regression, Statistical Hypothesis Generation and Testing, Python Libraries: NumPy, Pandas, Matplotlib

UNIT-II **(10 hours)**

Array manipulation using Numpy: Numpy array, Creating and various data types, indexing and slicing, swapping axes, transposing arrays, data processing using Numpy arrays.

UNIT-III **(15 hours)**

Data Manipulation using Pandas: Data Structures in Pandas: Series, Data Frame, 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 **(15 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.

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.
4. Chen D. Y, *Pandas for Everyone: Python Data Analysis*, Pearson,2018.
5. Miller J.D. *Statistics for Data Science,* Packt Publishing, 2017.

Practicals :

Use data set of your choice from Open Data Portal (https:// data.gov.in/) for the following exercises.

1. List of Practical based on NumPy
2. List of Practical based on Pandas
3. List of Practical based on Data Loading, Storage and File Formats
4. List of Practical based on Data Cleaning and Preparation
5. List of Practical based on DataWrangling
6. List of Practical based on Data Visualization usingmatplotlib

Use data set of your choice from Open Data Portal (https://data.gov.in/) for the following exercises.

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

Practice Questions sample

1. Load a Pandas data frame from a database.tdentify and count the missing values in adata frame. 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 to

i. Computemean,mode,median,standarddeviation,confidenceintervaland standard error for eachfeature
ii. Compute correlation between length and width of sepalfeature
iii. Find covariance between length of sepal andpetal
iv. Build contingency table for classfeature

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

DSE – 14
Data Science and Analytics 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		
Data Science and Analytics using Python	**4**	**3**	**0**	**1**	**Class XII pass with Mathematics**	DSC-01(Programming using Python)

Learning Objectives:

1. *To introduce the students to real-world data analysis problems.*
2. *To introduce students with concepts of data wrangling and aggregation.*
3. *To give students hands-on knowledge of Pandas.*

Learning Outcomes:

1. Use data analysis tools in the pandas library.
2. Load, clean, transform, merge and reshape data.
3. Create informative visualization and summarize data sets.
4. Analyse and manipulate time series data.

UNIT-I **(10 Hours)**

Introduction: Introduction to Data Science, Exploratory Data Analysis and Data Science Process. Motivation for using Python for Data Analysis, Introduction of Python shell iPython and Jupyter Notebook.

Essential Python Libraries: NumPy, pandas, matplotlib, SciPy, scikit-learn, statsmodels

UNIT-II **(10 Hours)**

Getting Started with Pandas: Arrays and vectorized computation, Introduction to pandas Data Structures, Essential Functionality, Summarizing and Computing Descriptive Statistics. Data Loading, Storage and File Formats.

Reading and Writing Data in Text Format, Web Scraping, Binary Data Formats, Interacting with Web APIs, Interacting with Databases Data Cleaning and Preparation. Handling Missing Data, Data Transformation, String Manipulation