22DS3T1: DATA SCIENCE

Course Name	Data Science		L	T	P	C	CIA	SEE	TM
Course Code	22DS3T1		4	0	0	4	30	70	100
Year of Introduction: 2021	Year of Offering: 2023	Year of Revision: 2023			Percentage of Revision: NA				
L-Lecture, T-Tutorial, **P**-Practical, **C**-Credits, **CIA**-Internal Marks, **SEE**-External Marks, **TM**-Total Marks									

Course Descriptive and Purpose: This course provides a comprehensive exploration of essential data manipulation and visualization tools, including NumPy, Pandas, Matplotlib, and techniques for handling, cleaning, merging, reducing, and transforming data.

Course Objectives: This course is designed to illustrate Modules of NumPy and Pandas, Matplotlib, Data, Universal Data Structures, Data Visualization, Data Cleaning, Data Fusion and Data Integration, Data Reduction, and Data Transformation and Massaging.

Specific objectives include:
- Overview of the Basic Functions of NumPy, Pandas and Matplotlib.
- Data, Universal Data Structure and Data Visualization.
- Data Pre-processing and Data Cleaning.
- Data Fusion and Data Integration and Data Reduction.
- Data Transformation and Massaging

Course outcomes:
Upon completion of the course
CO1: The course covers fundamental data manipulation and visualization techniques in Python, including NumPy and Pandas functions for data handling, Matplotlib for creating and customizing various plots, and techniques for subplots, resizing visuals, and saving them, providing a comprehensive foundation for data analysis and visualization in Python.
CO2: Acquire the knowledge and skills to distinguish between data types, perform data summarization and visualization, and investigate relationships between attributes, laying a solid foundation for data analytics and decision-making in real-world scenarios.
CO3: Proficient in conducting thorough data cleaning, enabling them to prepare high-quality datasets for analysis, make informed decisions, and build reliable predictive models, contributing to more accurate and actionable insights in data-driven projects.
CO4: Navigate data fusion and integration challenges, resulting in the ability to merge disparate data sources, reduce data redundancy, and perform data reduction using various techniques, thus enhancing their capabilities in managing and analyzing complex datasets for meaningful insights.
CO5: Possess the capability to apply data transformation and massaging techniques effectively, allowing them to prepare and optimize data for analysis, extract valuable features, and enhance the quality and usability of datasets for better decision-making and modeling outcomes.

UNIT-I (12 Hours)

Review of the Core Modules of NumPy and Pandas:
Overview of the Basic Functions of NumPy: The np.arange() function,The np.zeros() and np.ones() functions,The np.linspace() function
Overview of Pandas: (a)Pandas data access (b) Boolean masking for filtering a Data Frame (c) Pandas functions for exploring a DataFrame (d) Pandas applying a function I The Pandas groupby function (f) Pandas multi-level indexing (g) Pandas pivot and melt functions.

Review of Matplotlib:
1. Drawing the Main Plots in Matplotlib: (a) Summarizing numerical attributes using histograms or boxplots (b) Observing trends in the data using a line plot (c) Relating two numerical attributes using a scatterplot.
2. Modifying the Visuals: (a) Adding a title to visuals and labels to the axis (b) adding legends (c) Modifying ticks (d) Modifying markers
3. Subplots, Resizing Visuals and Saving them: (a) Resizing (b) Saving

UNIT-II (12 Hours)

Data: (a) What is Data? B) DIKW Pyramid (c) Data Preprocessing for Data Analytics versus Data Preprocessing for Machine Learning
The Most Universal Data Structure: A Table: (a) Data Objects b) Data Attributes
Types of Data Values: (a) **Analytics** Standpoint (b) Programming Standpoint
Information versus Pattern: (a) Understanding everyday use of the word "Information" (b) Statistical use of the word "Information" (c) Statistical meaning of the word "Pattern".
Analytic Goals:
Data Visualization-T, **Summarizing a Population: (**a) Example of summarizing Numerical Attributes (b) Example of summarizing Categorical Attributes.
Comparing Populations: (a) Example of comparing populations using Box Plots (b) Example of comparing populations using Histograms c) Example of comparing populations using Bar Charts.
Investigating the relationship between Two Attributes: (a) Visualizing the relationship between two Numerical Attributes (b) Visualizing the relationship between two Categorical Attributes (c) Visualizing the relationship between a Numerical Attribute and a Categorical Attribute.

UNIT-III (12 Hours)

The Pre-processing:
Data Cleaning Level I : Cleaning up the Table
The Levels, Pools, and Purposes of Data Cleaning: (a) Purpose of Data Analytics (b) Tools for Data Analytics (c) Levels of Data Cleaning (d) Mapping the purposes and tools of analytics to the levels of Data Cleaning.
Example 1: Unwise data collection **Example 2:** Reindexing (Multi-level Indexing), **Example 3:** Intuitive but long column titles.

Data Cleaning Level II: Unpacking, Restructuring, and Reformulating the Table
Example 1: Unpacking columns and reformulating the table**:** (a) Unpacking File Name (b) Unpacking Content I Reformulating a new table for Visualization (d) The last step – drawing the Visualization
Example 2: Restructuring the table.
Example 3: Level I and II Data Cleaning: Doing the analytics Using linear regression to create a predictive model.

Data Cleaning Level III: Missing Values, Outliers, and Errors
Missing Values: (a) Detecting Missing Values (b) Example of detecting Missing Values (c) Causes of Missing Values (d) Types of Missing Values I Diagnosis of Missing Values (f) Dealing with Missing Values.
Outliers: (a) Detecting Outliers (b) Dealing with Outliers.
Errors: (a) Types of Errors b) Dealing with Errors b) Detecting Systematic Errors.

UNIT- IV (12 Hours)

Data Fusion and Data Integration:
What are data fusion and data integration?
(a) Data Fusion versus Data Integration (b) Directions of Data Integration

F**requent challenges regarding Data Fusion and Integration:** (a) Challenge 1-Entity identification with Example (b) Challenge 2-Unwise data collection with Example (b) Challenge 3-Index mismatched formatting

with Example (c) Challenge 4 – Aggregation mismatch with example (d) Challenge 5- Duplicate data objects with example

Data Reduction:

The distinction between data reduction and data redundancy, The objectives of data reduction, Types of data reduction. **Performing Numerosity Data Reduction:** (a) Random Sampling (b) Stratified Sampling (c) Random over/under Sampling.

Performing dimensionality data reduction: (a) Linear Regression as a Dimension Reduction Method (b) Using a Decision Tree as a Dimension Reduction Method (c) Using a Random Forest as a Dimension Reduction Method.

UNIT-V (12 Hours)

Data Transformation and Massaging:

The whys of data transformation and massaging: (a) Data Transformation versus Data Massaging
(b) Normalization and Standardization

Binary Coding, Ranking Transformation, and Discretization: (a) Binary Coding of Nominal Attribute, Binary Coding or Ranking Transformation of Ordinal Attributes, Discretization of Numerical Attributes.

Attribute Construction: Construct one transformed Attribute from two Attributes.

Feature Extraction: Extract three Attributes from one Attribute.

Smoothing, Aggregation, And Binning: Smoothing, Aggregation, Binning.

Case Studies: 1.Mental Health in Tech, 2.Predicting COVID-19 Hospitalizations.

Text Books:

1. Hands-On Data Preprocessing in Python, Roy Jafari, Packt Publishing, 2022.
2. Python Data Analysis, Third Edition, Avinash Navlani Armando Fandango Ivan Idris, Packt Publishing, 2021.
3. Data Cleaning, Ihab F. Ilyas Xu Chu, Association for Computing Machinery, 2019.

PARVATHANENI BRAHMAYYA SIDDHARTHA COLLEGE OF ARTS & SCIENCE

(An Autonomous College in the jurisdiction of Krishna University)

M.Sc.(Computer Science), Third Semester

Course Name: Data Science C**ourse Code:** 22DS3T1

(w.e.f admitted batch 2022-23)

SECTION-A

Time: 3 Hours **Max. Marks: 70**

Answer ALL questions **(5×4 = 20 Marks)**

1. (a) Explain Slicing and Pandas Series within the Context of a Data Frame (CO1,L2)

(or)

(b) Explain the concept of Multi-Level Indexing in Pandas. (CO1,L2)

2. (a) What are the different types of data values from two different stand points (CO2,L2)

(or)

(b) How Data Visualization can be used to summarize numerical and categorical attributes with examples?.(CO2,L2)

3. (a) What are the different approaches used in dealing with Missing Values? CO3,L1)

(or)

(b) Define Outlier. What are the different tools used to detect outliers? (CO3,L1)

4. (a) Comparison between Data Fusion and Data Integration? (CO4.L2)

(or)

(b) Comparison between Data Reduction and Data Redundancy (CO4,L2)

5. (a) Comparison between Data Transformation and Data Massaging (CO5,L2)

(or)

(b) Demonstrate the creation of One Transformed Attribute Derived from two existing attributes. (CO5,L2)

SECTION-B

Answer Five Questions Choosing One Question from Each Unit.

All Questions Carry Equal Marks. **(5×10 = 50 Marks)**

6. (a) Explain the purpose and usage of basic NumPy Functions.(CO1, L2)

(or)

(b) Explain main types of Matplotlib Plots used in Exploratory Data Analysis such as Histograms Line Plots and Scatter Plots. (CO1,L2)

7. (a) Define the concept of DIKW Pyramid and compare Data Preprocessing techniques for Data Analytics and Machine Learning. (CO2,L1)

(or)

(b) How to investigate the relationship between Two Attributes in Data Visualization with Examples ? (CO2, L1)

8. (a) How to clean up a table at Level I of data cleaning with examples? (CO3,L1)

(or)

(b) How to Unpack, Restructure, and Reformulate the table at Level II of Data Cleaning with examples ? (CO3,L1)

9.(a) What are the frequent challenges regarding Data Fusion and Integration (CO4,L1)

(or)

(b) What are the different methods used to perform Numerosity Data Reduction. (CO4,L1)

10. (a) Explain the process of binary coding for Nominal Attributes, Ranking Transformation for ordinal attributes, and discretization for Numerical Attributes. (CO5,L5)

(or)

(b) Explain about Functional Smoothing and Rolling Smoothing with example. (CO5,L5)