# 20CS41O3 - FOUNDATIONS OF DATA SCIENCE

|  |  |  |  |
| --- | --- | --- | --- |
| Course Category: | Open Elective | Credits: | 3 |
| Course Type: | Theory | Lecture-Tutorial-Practical: | 3-0-0 |
| Prerequisite: | Require Transformation Techniques, Linear Algebra and Python Programming.  | Sessional Evaluation:Univ. Exam Evaluation:Total Marks: | 4060100 |
| Objectives: | * To impart knowledge on basics of data science, data manipulation and exploratory data analysis concepts that is vital for data science.
* To develop skills for applying tools and techniques to analyze, visualize and interpret data.
 |

|  |  |
| --- | --- |
| Course Outcomes | Upon successful completion of the course, the students will be able to: |
| CO1 | Demonstrate knowledge on the concepts of data science to perform mathematical computations using efficient storage and data handling methods in NumPy. |
| CO2 | Apply Data Preparation and Exploration methods using Pandas to perform data manipulation |
| CO3 | To determine Data transformation, String manipulation techniques |
| CO4 | To recognize Combining and merging datasets |
| CO5 | Create data visualization using charts, plots and histograms to identify trends, patterns and outliers in data using Matplotlib and Seaborn. |
| CO6 | Construct methods to analyze and interpret time series data to extract meaningful statistics |
| Course Content | UNIT-I**Introduction to Data Science:** Basic terminologies of data science, Types of data, Five steps of data science, Arrays and vectorized computation using NumPy - The NumPy ndarray: A multidimensional array object, Universal functions: Fast element-wise Array functions, Array-oriented Programming with arrays, File input and output with arrays, Linear algebra, pseudorandom number generation.UNIT-II**Data Exploration with Pandas:** Process of exploring data, Pandas data structures – Series, Data frame, Index objects; Essential functionality, Summarizing and computing descriptive statistics - Correlation and covariance, Unique values, Value counts and membership; Data loading, Storage, and file formats - Reading and writing data in text format, Binary data formats, Interacting with web APIs, Interacting with databases.UNIT-III**Data Cleaning, Preparation:** Handling missing data, Data transformation, String manipulation - String object methods, Regular expressions, Vectorized string functions in Pandas;UNIT-IV**Data Wrangling:** Data wrangling: join, Combine and reshape - Hierarchical indexing, Combining and merging datasets, Reshaping and pivoting.UNIT-V**Data Visualization with Matplotlib**: Plotting and visualization- A brief matplotlib API primer, Plotting with Pandas and Seaborn, Other python visualization tools; Data aggregation and Group operations- GroupBy mechanics, Data aggregation, Apply: General split-apply-combine, Pivot tables and Cross-tabulation.UNIT-VI**Time Series Analysis**: Date and time data types and tools, Time series basics, Date ranges, Frequencies, and shifting. Time zone handling, Periods and period arithmetic, Resampling and frequency Conversion – Downsampling, upsampling and interpolation, Resampling with periods; Moving window functions. |
| Text Books &ReferencesBooks | **TEXT BOOKS:**1. Wes McKinney, Python for Data Analysis, O ‘Reilly, 2nd Edition, 2017.

**REFERENCE BOOKS:**1. Sinan Ozdemir, Principles of Data Science, Packt Publishers, 2nd Edition, 2018.
2. Rachel Schutt, Cathy O‘Neil,Doing Data Science: Straight Talk from the Frontline, O‘Reilly, 2014.
 |
| E-Resources | 1. <https://swayam.gov.in/nd1_noc19_cs60/preview>
2. <https://towardsdatascience.com/>
3. <https://www.w3schools.com/datascience/>
4. <https://github.com/jakevdp/PythonDataScienceHandbook>
5. <https://www.kaggle.com>
 |