COURSE CONTENT

DATA HANDLING AND VISUALIZATION LABORATORY								
IV Semester: CSE (DS)								
Course Code	**Category**	**Hours / Week**			**Credits**	**Maximum Marks**		
ACDD04	Core	**L**	**T**	**P**	**C**	**CIA**	**SEE**	**Total**
		0	0	2	1	40	60	100
Contact Classes: Nil	**Tutorial Classes: Nil**	**Practical Classes: 45**				**Total Classes: 45**		
Prerequisites: Python Programming								

I. COURSE OVERVIEW:

Data handling is the process of collecting, organizing, and presenting the data in a way to analyze, make predictions, draw conclusions, and make decisions. Data visualization is a part of exploratory data analysis, a prior step before a full-pledged data analysis. This laboratory course is intended to offer practical knowledge and skills in both data handling and visualization. In this laboratory python packages such as NumPy, SciPy and Pandas for computations, and the visualization packages such as seaborn and matplotlib. Hands-on exercises are designed to explore the basic data importing, exploration, visualization, preliminary data analysis and data exporting techniques using core python and its packages. The expertise gained in this laboratory lays foundation for detailed data analysis that involves data modelling, analysis, evaluation and mining in scientific and engineering domains.

II. COURSE OBJECTIVES

The students will try to learn:

I. Installation and usage of python packages useful for data exploration and visualization.
II. Data handling using python in practice.
III. The practical knowledge of data visualization capabilities of python packages.

III. COURSE OUTCOMES

At the end of the course students should be able to:

CO 1 Tabulate the data from the CSV, XLS, TXT and JSON files as dataframes and export the dataframe to files.

CO 2 Make use of imputation techniques for wrangling the data using pandas package.

CO 3 Manipulate the python dataframes to form pivot tables and contingency tables.

CO 4 Manipulate the tabular data by joining multiple dataframes using pandas package.

CO 5 Explore the data using the data visualization techniques in python environment.

CO 6 Analyze the data for outliers to data trimming the data required for an authentic data analysis in python environment.

IV. COURSE CONTENT:

WEEK 1 Installation of python and related packages

1. Install python, and packages; NumPy, SciPy and Panda.
2. Study matrix operations: rank, inverse, condition number
3. Solving for simultaneous equations in 3 or 4 variables.

WEEK 2 Working with CSV files and XLS files.
1. Save a List to CSV, XLSX and TXT files.
2. Save a Dictionary to CSV, XLSX and TXT files.
3. Load data from CSV, XLSX and TXT pandas to a List.
4. Load data from CSV, XLSX and TXT pandas to a Dictionary.

WEEK 3 Basic operations on Dataframe.
1. Attribute filtering based on conditions.
2. Attribute filtering based on slicing.
3. Attribute filtering based on queries.

WEEK 4 Summary Statistics of the data
1. Compute ranking statistics of the data.
2. Compute statistical averages of numerical attributes.
3. Compute statistical ratios of numerical attributes.
4. Interpret the results.

WEEK 5 Handling Missing Values
1. Drop the rows containing missing values
2. Impute missing values with statistical averages.
3. Impute missing values using linear interpolation.
4. Interpret the results.

WEEK 6 Handling Time series data.
1. Display the date and time information in different formats.
2. Generate summary statistics during a period.
3. Compute rolling mean and rolling std deviations and plot.

WEEK 7 Visualization of categorial data
1. Plot categorical data as vertical and horizontal bar charts and label it.
2. Plot categorical data as vertical grouped bar chart and label it.
3. Plot categorical data as vertical stacked bar chart and label it.
4. Interpret the results.

WEEK 8 Visualization of correlations.
1. Plot the pair wise scatter plots of numerical attributes
2. Identify the type of correlations.
3. Interpret the results.

WEEK 9 Visualization of distributions
1. Plot the histograms of numerical data.
2. Plot the counts of categorial data.
3. Plot the data distributions (or densities).
4. Interpret the results.

WEEK 10 Visualization using box-and-whisker plots.
1. Compute the rank statistics of numerical attributes.

2. Create the box-and-whisker plots of numerical attributes.
3. Interpret the results.

WEEK 11 Handling outliers in the data.

1. Identify the outliers using quartile method.
2. Identify the outliers using standard deviation method.
3. Compare the performance of two methods.
4. Remove outliers from the data.
5. Interpret the results.

WEEK 12 Working with Data Tables.

1. Joining the data tables.
2. Exercises on contingency tables
3. Exercises on grouping data.

WEEK 13 Data Scaling and Transformation.

1. Scaling the data using different python scalers.
2. Normalization as a special case of data scaling.
3. Data transformation using standardization.
4. Compare the results and interpret.

WEEK 14 Web Scraping.

1. Scraping a list of items from a website.
2. Scraping data from a table.
3. Scraping images from a website.
4. Scraping data with pagination.

V. TEXTBOOKS

1. R. Nageswara Rao, "Core Python Programming, 3ed: Covers fundamentals to advanced topics like OOPS, Exceptions, Data structures, Files, Threads, Net", Dreamtech press, 3rd edition, 2021.
2. Eric Jacqueline Kazil & Katharine Jarmul," Data Wrangling with Python", O'Reilly Media, Inc, 2016.

VI. REFERENCE BOOKS:

1. Dr. Tirthajyoti Sarkar, Shubhadeep," Data Wrangling with Python: Creating actionable data from raw sources", Packet Publishing Ltd, 2019.
2. Making sense of Data: A practical Guide to Exploratory Data Analysis and Data Mining, by Glenn J. Myatt.
3. Wes McKinney, "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython", O'Reilly, 2nd Edition, 2018.
4. Dr. John P. Hoffmann, "Principles of Data Management and Presentation", 1st edition, 2017.
5. Jake VanderPlas, "Python Data Science Handbook: Essential Tools for Working with Data", O'Reilly, 2017.
6. Y. Daniel Liang, "Introduction to Programming using Python", Pearson, 2012.

VIII. ELECTRONIC RESOURCES

1. https://www.dataquest.io/blog/sci-kit-learn-tutorial/

2. https://www.ibm.com/support/knowledgecenter/en/SS3RA7_sub/modeler_tutorial_ddita/modeler_tutorial_ddita-gentopic1.html
3. https://archive.ics.uci.edu/ml/datasets.php
4. https://www.edx.org/course/analyzing-data-with-python
5. http://math.ecnu.edu.cn/~lfzhou/seminar/[Joel_Grus]_Data_Science_from_Scratch_First_Princ.pdf
6. https://www.programmer-books.com/introducing-data-science-pdf/

VIII. MATERIALS ONLINE

1. Course template
2. Lab Manual