DATA WRANGLING WITH PYTHON

V Semester: CSE (DS)

Course Code	Category	Hours / Week			Credits	Maximum Marks		
		L	T	P	C	CIA	SEE	Total
ACDC05	Core	3	1	0	4	30	70	100
Contact Classes: 45	Tutorial Classes: 15	Practical Classes: Nil				Total Classes:60		

Prerequisites: Python Programming.

I. COURSE OVERVIEW:

Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis. This course describes the importing of data from CSV and PDF files, data clean-up tasks such as elimination of bad data, duplicates and outliers, and data conditioning steps such as normalization and standardization. The course also discusses the data exploration for correlations and associations, and for providing statistical summaries of the given data. Several data visualizations such as plots, charts, maps, tables are also discussed. Finally, the principles of web scraping, web crawlers and spiders are presented. The knowledge and skills gained in this course are prerequisites for full-fledged data analysis.

II. COURSE OBJECTIVES:

The students will try to learn:

I	The concept and importance of data wrangling using Python.
II	The data cleaning and formatting techniques using Python.
III	The working with Excel, PDF and with non-relational database not supported bySQL using python.
IV	The application of techniques suitable for Web mining applications.

III. COURSE OUTCOMES:

After successful completion of the course, students should be able to:

CO 1	**Outline** the concept of and the steps in data wrangling process and the python basics necessary for implementing the data wrangling.	Remember
CO 2	**Summerize** the parsing approaches of the Excel as well as PDF Files for devising techniques to deal with uncommon file types.	Understand
CO 3	**Distinguish** between MySQL/PostgreSQL and NoSQL for storing and acquiring of data to and from the relational and the non-relational databases respectively.	Analyze
CO 4	**Explain** the operations involved in formatting and cleaning the data using Python for subsequent data analysis.	Understand
CO 5	**Make use of** python libraries for identifying outliers and correlations in the data, and visualizing the same efficiently.	Apply
CO 6	**Choose** appropriate method of web scraping and crawling based on web site model for acquring and storing data from world web within python framework.	Apply

IV. SYLLABUS:

MODULE – I: INTRODUCTION TO DATA WRANGLING (09)

What Is Data Wrangling? Importance of Data Wrangling, how is Data Wrangling performed? Tasks of Data Wrangling, Data Wrangling Tools, Introduction to Python, Python Basics, Data Meant to Be Read by Machines, CSV Data, JSON Data, XML Data.

MODULE – II: WORKING WITH EXCEL FILES AND PDFS (09)

Installing Python Packages, Parsing Excel Files, Getting Started with Parsing, PDFs and Problem Solving in Python, Programmatic Approaches to PDF Parsing, Converting PDF to Text, Parsing PDFs Using pdf miner, Acquiring and Storing Data, Databases: A Brief Introduction-Relational Databases: MySQL and PostgreSQL, Non-Relational Databases: NoSQL, When to use a Simple File, Alternative Data Storage.

MODULE – III: DATA CLEANUP (09)

Why Clean Data? Data Cleanup Basics, Identifying Values for Data Cleanup, Formatting Data, Finding Outliers and Bad Data, Finding Duplicates, Fuzzy Matching, RegEx Matching.

Normalizing and Standardizing the Data, Saving the Data, determining suitable Data Cleanup, Scripting the Cleanup, Testing with New Data.

MODULE – IV: DATA EXPLORATION AND ANALYSIS (09)

Exploring Data, Importing Data, Exploring Table Functions, Joining Numerous Datasets, Identifying Correlations, Identifying Outliers, Creating Groupings, Analyzing Data - Separating and Focusing the Data, Presenting Data, Visualizing the Data, Charts, Time-Related Data, Maps, Interactives, Words, Images, Video, and Illustrations, Presentation Tools, Publishing the Data - Open-Source Platforms.

MODULE – V: WEB SCRAPING (09)

What to Scrape and How, analyzing a Web Page, Network/Timeline, interacting with JavaScript, In-Depth Analysis of a Page, Getting Pages, Reading a Web Page - Reading a Web Page with LXML and XPath, Advanced Web Scraping - Browser-Based Parsing, Screen Reading with Selenium, Screen Reading with Ghost.Py, Spidering the Web - Building a Spider with Scrapy, Crawling Whole Websites with Scrapy.

V. TEXTBOOKS:

1. Jacqueline Kazil& Katharine Jarmul," Data Wrangling with Python", O'Reilly MediaInc., 2016.

VI. REFERENCE BOOKS:

1. Dr. Tirthajyoti Sarkar, Shubhadeep," Data Wrangling with Python: Creating actionable data from raw sources", Packt Publishing Ltd., 2019.
2. Stefanie Molin," Hands-On Data Analysis with Pandas", Packt Publishing Ltd.,2019
3. Allan Visochek," Practical Data Wrangling", Packt Publishing Ltd., 2017
4. TyeRattenbury, Joseph M. Hellerstein, Jeffrey Heer, Sean Kandel, Connor Carreras," Principles of Data Wrangling: Practical Techniques for Data Preparation", O'Reilly Media Inc., 2017

VII. WEB REFERENCES:

1. http://www.gbv.de/dms/ilmenau/toc/827365454.PDF
2. https://www.udemy.com/course/data-wrangling-with-python/
3. http://www.openculture.com/free-online-data-science-courses
4. https://www.classcentral.com/course/dataanalysiswithpython-11177