INSTITUTE OF AERONAUTICAL ENGINEERING
(Autonomous)
Dundigal, Hyderabad - 500 043

DEFINITIONS AND TERMINOLOGY

Department	**CSE(Data Science)**				
Course Title	**Data Wrangling with Python**				
Course Code	ACDC05				
Program	B. Tech				
Semester	V				
Course Type	CORE				
Regulation	UG20				
Course Structure	Theory			Practical	
	Lecture	Tutorials	Credits	Laboratory	Credits
	3	1	4	-	-
Course Coordinator	Ms J Sirisha, Assistant Professor, CSE (Data Science)				

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 by SQL using python.
IV	The application of techniques suitable for Web mining applications.

COURSE OUTCOMES

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

CO 1	**Outline** the concept of 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

DEFINITION AND TERMINOLOGY:

S.No	DEFINITION	CO's
	MODULE I	
	INTRODUCTION TO DATA WRANGLING	
1	**What is PIP?**	CO 1
	PIP is the package installer for Python. It is a command-line tool used to manage shared Python code and libraries.	
2	**How to determine the largest possible integer on your system?**	CO 1
	By using Python code: ⋙ import sys ⋙ sys.maxint	
3	**What is the output of the python instruction: pow(2,3,5)?**	CO 1
	The output is the remainder of the value (2 raised to 3) when modulo-divided by 5. i.e. 3	
4	**What are the data structures available in python?**	CO 1
	The builtin data structures are: lists, tuples, dictionaries, strings, sets and frozensets.	
5	**What is the mutability of a data structure in python?**	CO 1
	Mutability means that one can change an item present in the given data structure.	
6	**Differentiate between mutable and immutable data structures in python.**	CO 1
	Tuples and strings are immutables. Lists and Sets are mutables, whereas frozensets are immutable.	
7	**What is the output of the following python code?** **A = 10.8;B = 5;** **Result = 3 + B*14 +A // 3**	CO 1
	Result = 3 + 5*14 + 10.8 // 3 = 73 + 3.0 = 76.0	
8	**Differentiate between a module and a package in python.**	CO 1
	A module is a file containing Python definitions and statements. A package is a collection of modules. Modules that are related to each other are mainly put in the same package. When a module from an external package is required in a program, that package can be imported and its modules can be put to use.	
9	**Remove the multiplicates list Lst=[1, 2, 3, 1, 2, 19, 7, 7, 7],**	CO 1
	The set operation collects only the unique elemnets from the given list, hence the result is {1, 2, 3, 7, 19}	

10	**What is lambda() function in pyhton?**	CO 1
	The lambda() function can have one expression (only), whose evaluation and returning takes place. Moreover, this function can have any number of arguments. Example: def increment(n): return lambda x: x + n	
11	**What is the difference between append() and extend() methods in pyhton?**	CO 1
	The append() appends an object to the end of the list (e.g., another list) while the later appends each element of the iterable object (e.g., another list) to the end of the list. Example: Given X = ['a','b']; X.append(['c', 'f']) results in X = ['a','b',['c', 'f']] X.extend(['c', 'f']) results in X = ['a','b','c', 'f']	
12	**What is data wrangling?**	CO 1
	Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis.	
13	**What is the purpose of *agate* package in python?**	CO 1
	It's a data-analysis library built by Christopher Groskopf. It will allow us to start looking at some basic features of the given data such as descriptive statistics, outlier removal,	
14	**Give the list: a = [1,2,3,4], the pyhton instruction: b = (2*x for x in a) results in the new list b. What is b?**	CO 1
	b=[2,4, 6, 8]	
15	**What is data cleanup?**	CO 1
	Data cleanup is the standardization of the data format to improve quality and consistency. This is done by fixing structural errors and typos, changing null values, removing outliers and so on.	
16	**What is Data Exploration and Visualization?**	CO 1
	Data wrangling often involves exploratory data analysis (EDA) and visualization to understand the characteristics of the data, identify trends, patterns, and correlations, and guide further analysis.	
17	**What is Data Imputation?**	CO 1
	Data wrangling may require imputing missing values, where the values are estimated or predicted based on other available data, ensuring that the dataset remains complete.	
18	**What is Data Reshaping?**	CO 1
	Data reshaping tasks involve changing the structure of data, such as pivoting, melting, or transposing, to make it suitable for analysis or visualization.	
19	**What is Data Labeling?**	CO 1
	Data labeling involves adding descriptive labels or categories to data points, making them more interpretable or suitable for classification tasks.	

20	**What is Data Splitting?** Data splitting is the process of dividing the dataset into training, validation, and testing sets for model development and evaluation.	CO 1
MODULE II		
WORKING WITH EXCEL FILES AND PDFS		
1	**What does JSON stands for?** JSON stands for JavaScript Object Notation, which is an open data interchange format that is both human and machine-readable. JSON is independent of any programming language.	CO 1
2	**What is a CSV file?** CSV files are file formats that contain plain text values separated by commas. CSV files can be opened by any spreadsheet program: Microsoft Excel, Open Office, Google Sheets, or any text editor.	CO 1
3	**What is a delimitor in a CSV data file?** A delimiter is a sequence of one or more characters for specifying the boundary between separate, independent regions in plain text, mathematical expressions or other data streams. In CSV file the default delimiter comma (,) is used to separate fields.	CO 1
4	**What is an xls or xlsx file?** XLS files are Microsoft Excel's workbook files in use between 97-2003. Later Excel versions use the XLSX extension.	CO 1
5	**What is spyder?** Spyder is an open-source cross-platform IDE. The Python Spyder IDE is written completely in Python. It is designed by scientists and is exclusively for scientists, data analysts, and engineers.	CO 2
6	**How data is stored in a DSV file?** A delimited text file is a text file used to store data, in which each line represents a single record. In delimiter-separated values (DSV) file data is stored in two-dimensional arrays of data by separating the values in each row with specific delimiter characters (e.g. comma, tab, semicolon, etc).	CO 2
7	**What is the role of PDF interpreter in PDF file parsing?** A delimited text file is a text file used to store data, in which each line represents a single record. characters (e.g. comma, tab, semicolon, etc).	CO 2
8	**What is pdfminer?** PDFMiner is a tool for extracting information from PDF documents. Unlike other PDF-related tools, it focuses entirely on getting and analyzing text data. PDFMiner allows one to obtain the exact location of text in a page, as well as other information such as fonts or lines.	CO 2

9	**What is qualifier?**	CO 2
	If a field contains the delimiter character within its text, the program interprets this as the end of the field rather than as part of the text. In order to prevent this, it must have a qualifier. The qualifier is placed around each field to signify that it is the same field. The most common qualifier is double quotes (”).	
10	**What is PDFResourceManager?**	CO 2
	PDFResourceManager is one of the classes available with pdfmailer. It is used to store shared resources such as fonts or images.	
11	**What are the two classes to parse a PDF file using pdfminer?**	CO 2
	To parse PDF files, we need at least two classes: PDFParser and PDFDocument. PDFParser fetches data from a file, and PDFDocument stores it.	
12	**What is the data storage format of noSQL database?**	CO 2
	NoSQL like other nonrelational databases stores data in a flat format, usually JSON.	
13	**What is the popular NoSQL database to which python can connect?**	CO 2
	MongoDB is the most popular NoSQL database framework used by python application to connect to. First we need to install the drivers and then use Python to connect.	
14	**What is relational data?**	CO 2
	Relational data usually uses a series of unique identifiers to actively match datasets. In SQL, we normally call them IDs. These IDs can be used by other sets of data to find and match connections.	
15	**What is Psycopg?**	CO 2
	Psycopg is the most popular PostgreSQL database adapter for the Python programming language. Its main features are the complete implementation of the Python DB API 2.0 specification and the thread safety.	
16	**What does OCR stands for?**	CO 2
	OCR (Optical Character Recognition) is a technology that converts scanned or image-based PDFs into machine-readable text. OCR is used when the PDF content is not natively selectable.	
17	**What is Page Navigation?**	CO 2
	PDF parsing tools may offer features to navigate between pages, locate specific content, and extract data from particular sections of a PDF document.	
18	**What is PDF Annotations?**	CO 2
	Annotations in PDFs, like comments, highlights, and form field data, can be programmatically accessed and manipulated.	

19	**What is Data Integration?** Extracted data from PDFs can be integrated into databases, spreadsheets, or other data storage and analysis tools.	CO 2
20	**What is Metadata Extraction?** Metadata, such as document title, author, creation date, and keywords, can be extracted programmatically to provide additional information about the PDF document.	CO 2
	MODULE III	
	DATA CLEANUP	
1	**What is the purpose of replacing column headers in an excel file?** Some headers are very short and not understadnable. So they must be replaced with longer English ones for better understanding.	CO 4
2	**What is the zip() method in python?** Python zip() method takes iterable or containers and returns a single iterator object, having mapped values from all the containers. It is used to map the similar index of multiple containers so that they can be used just using a single entity.	CO 4
3	**What does an 'r' before a string tells in python?** An 'r' before a string tells the Python interpreter to treat backslashes as a literal (raw) character. Normally, Python uses backslashes as escape characters.	CO 4
4	**What is the purpose of *isspace()* method of python's string class?** The Python string class's isspace method returns True if the string contains only. spaces.	CO 4
5	**What is the use of *find()* method of python's string class?** The string's *find()* method returns the index of the first match. If it finds no match in the string, it returns -1. This code tests for both / and :, which are commonly used in time strings.	CO 4
6	**What is the use of *fillna()* method of python?** The *fillna()* method is used to fill NaN/NA values on a specified column or on an entire DataaFrame with any given value.	CO 4
7	**What is the simplest method to eliminate the duplicates in a python's list?** The duplicates in a python's list can be eliminated by implicit type conversion of list to set.	CO 4
8	**What are the methods used to drop rows having a missed entry in a pandas dataframe?** First the missed values are replaced with NaN using *fillna()* and then the rows with NaN are dropped using the method *dropna()*.	CO 4

9	**What is data cleansing?** data cleaning or data scrubbing, is the process of fixing incorrect, incomplete, duplicate or otherwise erroneous data in a data set.	CO 4
10	**What is Delimiter?** A delimiter is a character or sequence used to separate or distinguish data fields in a structured format, such as a comma in CSV files or a tab in tab-delimited files.	CO 4
11	**What is Metadata?** Metadata is data about data. It includes information about the source, structure, and meaning of data elements, facilitating data management and interpretation.	CO 4
12	**What is Outliers?** Outliers are data points that deviate significantly from the rest of the data in a dataset.	CO 4
13	**What is Box Plot?** A box plot, also known as a box-and-whisker plot, is a graphical representation used to identify outliers in a dataset.	CO 4
14	**What is IQR (Interquartile Range)?** The IQR is a measure of statistical dispersion. It is used in conjunction with box plots to identify outliers.	CO 4
15	**What is Fuzzy Matching?** Fuzzy matching is a technique used to identify duplicates that are not exact matches.	CO 4
16	**What is Record Linkage?** Record linkage is the process of identifying and connecting records that refer to the same entity or individual across different data sources.	CO 4
17	**What is Threshold Tuning?** Threshold tuning is the process of finding the optimal similarity score threshold for a specific fuzzy matching task, balancing precision and recall.	CO 4
18	**What is De-duplication ?** De-duplication is the action of removing or consolidating duplicate records in a dataset to maintain data quality and integrity.	CO 4
19	**What is Scoring or Weighting ?** Scoring or weighting is a method to assign values to data fields based on their importance in identifying duplicates. Fields with higher scores contribute more to the matching process.	CO 4
20	**What is Blocking Key?** A blocking key is a field or attribute used to divide data into blocks for more efficient duplicate detection. Records within the same block are compared to identify potential duplicates.	CO 4

MODULE IV		
DATA EXPLORATION AND ANALYSIS		
1	**what is Metacharacters?** Metacharacters are characters in a regular expression that have special meanings, such as . (dot) for any character, * (asterisk) for zero or more repetitions, and + (plus) for one or more repetitions.	CO 4
2	**what is Greedy vs. Non-Greedy Matching:?** Greedy matching attempts to match as much text as possible, while non-greedy (lazy) matching attempts to match as little as possible. This behavior is controlled by quantifiers, such as * and *?.	CO 4
3	**what is Min-Max Scaling?** Min-max scaling, also known as feature scaling, rescales data to a specific range, usually between 0 and 1.	CO 4
4	**what is Data Exploration?** Data exploration is the process of summarizing, visualizing, and understanding the main features and properties of a dataset. It includes various tasks and techniques to uncover insights.	CO 5
5	**what is Central Tendency?** Central tendency refers to the tendency of data to cluster around a central value. Measures of central tendency include the mean (average), median (middle value), and mode (most frequent value).	CO 5
6	**what is Scatter Plot?** A scatter plot is a graphical representation that displays the relationship between two numerical variables as points on a two-dimensional plane.	CO 5
7	**what is Data Importing ?** Data importing is the process of bringing external data into a software or tool for analysis. This can involve reading data from files (e.g., CSV, Excel, JSON) or connecting to databases, APIs, or web sources.	CO 5
8	**what is ETL ?** ETL(Extract, Transform, Load) is a data integration process involving the extraction of data from multiple sources, transforming it to fit a common data model, and loading it into a destination, such as a data warehouse.	CO 5
9	**what is Table Functions?** Table functions are built-in or user-defined functions that operate on tables in a database. They allow you to perform various operations, such as filtering, aggregation, and joining.	CO 5
10	**what is Correlation Analysis?** Correlation analysis is a statistical technique used to assess the strength and direction of the relationship between two or more variables. It is measured using correlation coefficients like Pearson's r.	CO 5
11	**what is Outlier Detection?** Outlier detection is the process of identifying data points that significantly deviate from the majority of data points in a dataset.	CO 5

12	**What is Data Grouping?** Data grouping involves categorizing data into groups or clusters based on shared characteristics, attributes, or criteria. It's commonly used for summarization and analysis.	CO 5
13	**What is Data Repository?** A data repository is a digital platform or archive where datasets are stored, managed, and made accessible to the public. Well-known examples include data.gov, Zenodo, and the Harvard Dataverse.	CO 5
14	**What is Data Interoperability?** Data interoperability involves the ability of datasets published on open-source platforms to work seamlessly with various software tools, platforms, and data standards.	CO 5
15	**What is Data Data Collaboration?** Open-source platforms often support collaboration among data contributors and users, enabling the crowd-sourcing of data, insights, and solutions.	CO 5
16	**What does API stands for?** API stands for Application Programming Interface.Some open-source platforms offer APIs that enable developers to programmatically access and integrate datasets into their applications and services.	CO 5
17	**What is Embedding?** Embedding is the process of including the presentation or parts of it within web pages, blogs, or other digital documents to make it viewable without leaving the page.	CO 5
18	**What is Data Segmentation?** Data segmentation involves dividing a dataset into smaller segments or groups based on common characteristics or attributes. This can help in focusing the analysis on specific sub populations.	CO 5
19	**What is Data Stratification?** Data stratification involves dividing data into distinct strata or layers based on specific characteristics. This helps in analyzing and comparing data within each stratum.	CO 5
20	**What is Data Segment Analysis?** Segment analysis involves conducting separate analyses on distinct segments or subsets of data, such as different customer segments or geographic regions.	CO 5
MODULE V		
WEB SCRAPING		
1	**What is XPath Functions?** XPath includes functions for performing operations on nodes and values, such as text(), contains(), and substring().	CO 6

2	**What is Headless Browsers?** Headless browsers are web browsers that operate without a graphical user interface (GUI). They can be controlled programmatically to navigate websites, interact with page elements, and extract data.	CO 6
3	**What is Selenium ?** Selenium is an open-source framework for automating web browsers. It provides a set of tools and libraries for controlling browser behavior programmatically.	CO 6
4	**What is WebDriver?** WebDriver is a component of Selenium that acts as a remote control for web browsers. It allows you to create automated scripts to interact with web pages.	CO 6
5	**What is XPath and CSS Selectors?** XPath and CSS selectors are patterns used to locate specific HTML elements on a web page. Selenium uses these selectors to target and interact with elements.	CO 6
6	**What is Scrapy Shell?** The Scrapy shell is an interactive environment that allows you to experiment with and test Scrapy selectors and commands before implementing them in your spider.	CO 6
7	**What is Selectors?** Selectors are patterns or expressions used to identify and locate specific elements within an HTML document, such as CSS selectors, XPath, or regular expressions.	CO 6
8	**What is Web Scraping?** Web scraping is the automated extraction of data from websites. It involves sending HTTP requests to web pages, parsing their HTML, and extracting specific information.	CO 6
9	**What is Scraping Targets?** Scraping targets are the specific pieces of data or elements on a web page that you want to extract. These can be product prices, news headlines, weather information, or any other relevant content.	CO 6
10	**What is Crawling?** Web crawling is the process of systematically navigating through a website by following links to access multiple pages. It's used to collect data from various parts of a website.	CO 6
11	**What is Robots.txt?** Robots.txt is a text file placed on a website to indicate which parts of the site are off-limits to web crawlers or scrapers. Compliance with robots.txt is essential to respect a website's terms of service.	CO 6

12	**What is JavaScript ?**	CO 6
	JavaScript is a programming language commonly used for adding interactivity and dynamic behavior to web pages. It is supported by most web browsers and can be executed directly within the browser.	
13	**What is DOM (Document Object Model) ?**	CO 6
	The DOM is a programming interface that represents the structure and content of a web page. It allows JavaScript to interact with and manipulate elements on the page.	
14	**What is Script Tag?**	CO 6
	In HTML, the ¡script¿ tag is used to include JavaScript code within a web page. JavaScript can be embedded directly in HTML or loaded from external files.	
15	**What is AJAX (Asynchronous JavaScript and XML) ?**	CO 6
	AJAX is a set of web development techniques that use JavaScript to send and receive data from a web server without requiring a full page refresh. It enables dynamic content updates.	
16	**What is LXML ?**	CO 6
	LXML is a Python library that provides a high-performance, easy-to-use interface for working with XML and HTML documents. It allows you to parse and manipulate structured data from web pages.	
17	**What is XML (eXtensible Markup Language) ?**	CO 6
	XML is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. It is often used for structured data exchange.	
18	**What is XPath Expression ?**	CO 6
	An XPath expression is a string that specifies the location of elements or nodes in an XML or HTML document. Expressions are used to select and query elements.	
19	**What is XPath Axes?**	CO 6
	Axes in XPath define the context for element selection and traversal. Common axes include the child axis (/), descendant axis (//), and attribute axis (@).	
20	**What is XPath Predicates?**	CO 6
	Predicates in XPath are conditions that filter elements based on specified criteria. They are used within square brackets, such as [contains(@class, 'example')].	

Prepared by
Ms J Sirisha, Assistant Professor

HOD,CSE(DS)