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

COMPUTER SCIENCE ENGINEERING (DATA SCIENCE)

QUESTION BANK

Course Title	**Data Wrangling with Python**				
Course Code	ACDC05				
Program	B.Tech				
Semester	V	CSE (Data Science)			
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 OBJ ECTIVES:

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

QUESTION BANK:

Q.No	QUESTION	Taxonomy	How does this subsume the level	CO's
MODULE I				
INTRODUCTION TO DATA WRANGLING				
PART A-PROBLEM SOLVING AND CRITICAL THINKING QUESTIONS				
1	Is it possible to completely eliminate errors or inconsistencies in data through the data wrangling process?	Apply	It is not possible to completely eliminate errors or inconsistencies in data through the data wrangling process. Data wrangling aims to minimize errors and inconsistencies by cleaning, transforming, and integrating data, but it cannot guarantee complete elimination.	CO 1
2	Create a numerical array of 20 elements using a numpy built-in-sequence, and then convert it into a 3-dimensional array. Explain the methods used in this process.	Apply	The learner will try to recall and understand the usage of numpy methods, apply the knowledge to create and reshape the array.	CO 1
3	Justify the role of pandas in data science applications.	Apply	The learner will try to recall and understand the pandas dataframes and methods, apply the knowledge to cleanse and explore the data.	CO 1

4	Can data wrangling be performed effectively without a deep understanding of the underlying data domain?	Apply	Data wrangling can be performed more effectively with a deep understanding of the underlying data domain. Having domain knowledge helps in understanding the context, identifying data quality issues, and making informed decisions during the wrangling process.	CO 1
5	In what ways can Python programming language enhance the efficiency and effectiveness of data wrangling processes?	Apply	Python programming language can enhance the efficiency and effectiveness of data wrangling processes in several ways	CO 1
6	Identify and discuss numpy methods useful for computing rank statistics and interquartile range with a numerical example.	Apply	The learner will try to recall numpy methods, and understand the definitions of rank statistics, and interquartile range, and then select the appropriate method(s).	CO 1
7	Develope python code to compute the median of a vector with (a). odd number of elements (b). even o number of elements, without using any python/ numpy/pandas methods. The code must contain the basic instructions of python.	Apply	The learner will try to recall numpy methods, and understand the definition of median, and then develope python code.	CO 1
8	Is it feasible to perform data wrangling on real-time streaming data, or is it more suitable for batch processing?	Apply	Performing data wrangling on real-time streaming data is feasible, but it presents additional challenges compared to batch processing. Real-time data wrangling requires handling data as it arrives in a continuous stream, often with limited time for processing and analysis.	CO1

9	Can the 'loc' function in pandas be used to modify data directly in the original DataFrame?	Understand	the 'loc' function in pandas is primarily used for label-based indexing and selection of data from a DataFrame. It is not intended for direct modification of data in the original DataFrame. To modify data, you would typically use assignment operators or other appropriate methods provided by pandas.	CO1
10	Create two tables and use table functions to merge/join tables with the help of python	Apply	Creating two tables and using table functions to merge/join tables with Python involves applying the "Apply" level of Bloom's Taxonomy, as it requires using knowledge and skills to perform a specific task	CO1
PART-B LONG ANSWER QUESTIONS				
1	Discuss in details, the benefits of Data Wrangling to business applications.	Understand	The learner will try to recall definition and tasks of data wrangling, and understand the same in the context of busineess needs.	CO 1
2	Distinguish how the data is stored in machine-readable data files.	Apply	The learner will try to recall the formats of CSV/ XLSX/ JSON files, understand and apply the knowledge to compare them.	CO 1
3	Discuss in details, the benefits of Data Wrangling in the data analytic applications?	Understand	The learner will try to recall definition and tasks of data wrangling, and identify the same as prerequisite steps for data analysis.	CO 1
4	Data validating and Data publishing are the two tasks of Data Wrangling. Discriminate them.	Understand	The learner will try to recall definition and tasks of data wrangling, compare both the given tasks and conclude.	CO 1

5	Explain in detail the Data Structuring process in the context of Data Wrangling?	Understand	The learner will try to recall the tasks of data wrangling, and the definition of data structuring, and then find the link between them.	CO 1
6	Explain in detail the popular data file formats: CSV, XSLX and JSON.	Understand	The learner will try to recall the formats of CSV, XSLX and JSON data files, then explain or comapre them.	CO 1
7	Illustrate the structure of a basic XML file.	Understand	The learner will try to recall the stucture of XML file, and discuss about its components and their purpose.	CO 1
8	Contrast a numpy array with a python list.	Understand	The learner will try to recall definitions of a numpy array and a python list, then compare their methods() and attributes.	CO 1
9	Explain how data enriching forms an important step in data wrangling process.	Understand	The learner will try to recall the tasks of data wrangling, and the definition of data enriching, and then identify a link between them.	CO 1
10	Discriminate between pandas series and pandas dataframe.	Understand	The learner will try to recall the definition of series and dataframe in pandas, and then identify the differences.	CO 1
11	Justify the Data preparation consumes 80% of data scientists' workload. Justify.	Compare	The learner will try to recall definition of data wrangling, understand and apply the wrangling tasks, and then compare the involved effort with that of other data science tasks.	CO 1
12	Justify that the data validation is an important step in data wrangling process.	Compare	The learner will try to recall and understand the definitions of data wrangling, and data validation, and compare both to justify the given statement.	CO 1

13	Compare and contrast different indexing methods of a pandas dataframe.	Apply	The learner will try to recall and understand the definitions of data wrangling, and data validation, and compare both to justify the given statement.	CO 1
14	Distniguish among the formats of atleast three data files, and discuss salient features of each file format.	understand	The learner will try to recall few file formats, and understand their structure and compare.	CO 1
15	Discuss about distinct features of XML and JSON file formats, in the context of data storage and meta data.	understand	The learner will try to recall the required file formats, understand their structure and discuss.	CO 1
16	How can I create a basic line plot using Matplotlib?give an example?	understand	How can I create a basic line plot using Matplotlib?How can I create a basic line plot using Matplotlib?	CO 1
17	Discriminate between pandas loc and pandas iloc.	Understand	The learner will try pandas, and then identify the differences.	CO 1
18	Distniguish among the formats of atleast three data files, and discuss salient features of each file format.	understand	The learner will try to recall few file formats, and understand their structure and compare.	CO 1
PART-C SHORT ANSWER QUESTIONS				
1	What is data wrangling? Explain in brief.	Understand	The learner will try to recall the definition of data wrangling and explain different tasks or steps involved.	CO 1
2	Explain the importance of data enriching?	Understand	The learner will try to recall the definition of data wrangling and data enriching, and then find its link between them.	CO 1
3	Briefly explain the Data discovery?	Understand	The learner will try to recall the definition of data discovery, and identify its application.	CO 1
4	What is numpy package?	Remember	-	CO 1

5	Give the Sample record of a JSON file.	Remember	-	CO 1
6	What is the structure of an XML file?	Remember	-	CO 1
7	Distinguish between. JSON and .CSV file formats.	Understand	The learner will try to recall the formats of JSON and .CSV files, and then tabulate the differences.	CO 1
8	Tabulate the precedence of operators in python in the descending order.	Remember	-	CO 1
9	What is pandas? Explain in brief.	Understand	-	CO 1
10	What is data parsing?	Remember	-	CO 1
11	What are the data structures available in pandas? Discuss in brief.	Remember	-	CO 1
12	What are the goals of data wrangling with ython?	Understand	-	CO 1
13	What are the key steps to data wrangling with python?	Understand	-	CO 1
14	how is Data Wrangling performed?	Understand	-	CO 1
15	What is the different between Datatypes and containers?	Remember	-	CO 1
16	What is the different between mutable Datatypes and imutable datatypes?Give an examples?	Remember	-	CO 1
17	Write about data wrangling tools?Give an examples?	Understand	-	CO 1
18	Briefly explain the data sets? how can we get the data sets?	Understand	-	CO 1
19	what is data munging? Explain briefly.	Understand	-	CO 1
20	How we can read the excel file? Give the syntax .	Understand	-	CO 1

MODULE II				
WORKING WITH EXCEL FILES AND PDFS				
PART-A PROBLEM SOLVING AND CRITICAL THINKING QUESTIONS				
1	Considering few public datasets available on world web, discuss their content and utility for exploring data science problems.	Understand	The learner will try to recall the public datasets, their content and identify their usage.	CO 3
2	Justify the role of two python objects involved in Python DB-API.	Analyze	The learner will try to recall the DB-API and its objects, understand the characteristics of objects, analyze and then justify their role.	CO 3
3	Illustrate the mechanism by which the Python program communicates with any DBM System.	Understood	The learner will try to recall DB-API, database connection and query protocols of a typical DBM system and then explain a typical communication session using pyhton.	CO 3
4	Highlight the distinct features of mySQL database compared to other databases, and the features of mySQL used for datbase transaction.	Apply	The learner will try to recall the operation of relational DBMS, and the SQL programming, and then summarize the mySQL system.	CO 3
5	Distinguish between SQLite and MySQL database engines by comparing their capabilities.	Apply	The learner will try to recall the features of SQLite and MySQL database systems (engines), their interface methods and then make the required comparison.	CO 3
6	Distinguish between PostgreSQL and NoSQL database system	Understood	The learner will try to recall the fundamentals of relational and nonrelational databases, and then compare their characteristics.	CO 3

7	Discuss the methodology through a flowchart for inserting a data object into a non-relational databse using MongoClient. Also provide python code.	Apply	The learner will try to recall the definiiton of a python data sequence object and the usage of pymongo package, and then develope the required algorithm and flowchart.	CO 3
8	Compare the PDF parsing methods using pdfminer and PyPDF2.	Apply	The learner will try to recall packages related to PDf file handling, understand their usage and compare the methods.	CO 2
9	Explain how to transfer the data from a JSON file to a CSV file. Develope the python code for the same.	Apply	The learner will try to recall packages related to JSON and CSV file handling, understand their usage and develope the algorithm and python code.	CO 2
10	Explain how to transfer the data from a CSV file to a Excel file. Develope the python code for the same.	Apply	The learner will try to recall packages related to Excel and CSV file handling, understand their usage and develope the algorithm and python code.	CO 2
PART-B LONG ANSWER QUESTIONS				
1	Write the algorithmic steps in transferring data from pandas dataframe to a CSV file, along with python code.	Apply	The learner will try to recall the definition of dataframe and the package related to file handling and develope the required algorithm.	CO 2
2	What is a delimiter in the context of data files? Discuss the procedure to write a python nested list object to an XLS file without using pandas.	Apply	The learner will try to recall the definition of nested list and the package related to file handling and develope the required algorithm.	CO 2
3	Explain the algorithm to create a dictionary object in python and write it as an XLSX file using pandas. Write the python code also.	Apply	The learner will try to recall the definition of dictionary and dataframe, and the package related to file handling and develope the required algorithm.	CO 2

4	Explain the procedure to create a nested list in python and write it as a XLS file without using pandas. Write the python code also.	Apply	The learner will try to recall the definition of a nested list and the package related to file handling and develope the required code.	CO 2
5	How do you read data from CSV file without using pandas? Write down the python code also.	Apply	The learner will try to recall the definition of dataframe and the package related to file handling and develope th erequired algorithm.	CO 2
6	How do you extract text data from a PDF file using pyPDF2 in python. Explain with an algorithm.	Apply	The learner will try to recall packages related to PDf file handling , and develope the required algorithm.	CO 2
7	How do you extract text data from a PDF file using pdfminer in python. Explain with an algorithm	Apply	The learner will try to recall packages related to PDF file handling , and develop the required algorithm.	CO 2
8	How do you write text data to a PDF file using python. Explain with an algorithm.	Apply	The learner will try to recall the package related to file handling and develop the required algorithm.	CO 2
9	How you extract text data from a PDF file using python without using pandas. Write down the python code also.	Apply	The learner will try to recall the package related to file handling and develop the required algorithm.	CO 2
10	How to read an XML file into a data object using python?	Apply	The learner will try to recall the format of an XML file and the package related to file handling and develop the required algorithm.	CO 2
11	Explain foundations and structure of relational databases?	Apply	The learner will try to recall the concept of relational databases.	CO 2
12	How do you establish a connection to a MySQL database using Python, and what are the essential libraries and modules involved?	Apply	The learner will try to recall the required steps to make connection between MySQL and Python.	CO 2

13	How to set up data into a relational database with MySQL. Explain with steps.	Apply	The learner will try to recall the required steps to set up data.	CO 2
14	How to populate data with MySQL. Explain with steps.	Apply	The learner will try to recall the required steps to populate the data.	CO 2
15	How to install psycopg2 and establish a connection PostgreSQL with Python.	Apply	The learner will try to recall the connection establishment with Python.	CO 2
16	Explain the advantages and disadvantages of NoSQL databases compared to traditional relational databases?.	Apply	The learner will try to recall the concept of NoSQL usage in realtime.	CO 2
17	How to connect to NoSQL databases using Python (with MongoDB).	Apply	The learner will try to recall the concept of MongoDB.	CO 2
18	how to create a connection and work with a SQLite database using the dataset library.	Apply	The learner will try to recall the concept of pip simple library dataset working with SQLite.	CO 2
19	Explain alternative data storage paths in detail.	Analyze	The learner will try to recall the concept of various data storage's.	CO 2
20	Explain dataset installation and usage with code.	Apply	The learner will try to recall the concept of dataset usage.	CO 2
PART-C SHORT ANSWER QUESTIONS				
1	How to read a CSV file into a pandas dataframe?	Remember	–	CO 2
2	How to read a JSON file into a pandas dataframe?	Remember	–	CO 2
3	How to read an XML file into a dataframe using python?	Remember	–	CO 2
4	How to create a XLSX file from a pandas dataframe?	Remember	–	CO 2
5	How to create a XLSX file from a pandas dataframe with the default delimiter?	Remember	–	CO 2
6	How to create a CSV file from a dataframe using a delimiter '—' in python?	Remember	–	CO 2

7	What are the elements of a PDF file?	Remember	–	CO 2
8	How to read a JSON file into a pandas dataframe?	Remember	–	CO 2
9	What is a separator in the context of a data file? Briefly discuss.	Understand	The learner will try to recall the CSV data file format and explain the role of seperator in a file	CO 2
10	How to read an XML file into a python data object?	Remember	–	CO 2
11	What are NoSQL databases?	Understood	The learner will try to recall the fundamentals of non-SQL databases and explain their distinctness.	CO 3
12	Differentiate between Horizontal and vertical scaling-up in reference to database servers.	Understood	The learner will try to recall the concept of Horizontal and vertical expansion of computing resources in general, and then explain the scalingup of database servers.	CO 3
13	Explain pdf elements?	Remember	The learner will try to recall the fundamentals of non-SQL databases and explain their distinctness. –	CO 2
14	what is the PDF Parsing?	Remember	The learner will try to recall the fundamentals of non-SQL databases and explain their distinctness. –	CO 2
15	what are relational databases?	Remember	The learner will try to recall the various relational databases. –	CO 2
16	what are non-reational databases?	Remember	The learner will try to recall the various non-relational databases. –	CO 2
17	Write a Python dictionary of the data to save?	Remember	The learner will try to recall the fundamentals of dictionaries creation. –	CO 2
18	Write SQL command to create a new table called ”datasources” ?	Remember	The learner will try to recall the fundamentals of SQL Queries. –	CO 2
19	How to showcase all of the data sources user stored in ”datasources” table?	Understand	The learner will try to recall the fundamentals of database queries. –	CO 2

20	How to insert first data source into a new table?	Remember	The learner will try to recall the fundamentals of table queries. –	CO 2
MODULE III				
DATA CLEANUP				
PART A-PROBLEM SOLVING AND CRITICAL THINKING QUESTIONS				
1	Justify why normalization is not preferable to standardization, in case of gaussian attributes.	Analyse	The learner will try to recall and understand the scaling methods:normalization and standardization, and then apply on features.	CO 4
2	An N-dimensional data sample can be represented as a point object in N-dimensional feature space. Justify with an example of a 2-dimensional space.	Apply	The learner will try to recall concept of vector spaces, and then apply visualization tool to view data samples.	CO 4
3	Justify why normalization is an appropriate method of scaling for non-gaussian attributes.	Analyse	The learner will try to recall and understand the scaling methods:normalization and standardization, and then apply on features.	CO 4
4	Explain the equations for computing the mean and RMS values of an array. Develope python for the same from the basic python instructions without using the built-in-methods.	Apply	The learner will try to recall and understand the scaling methods:normalization and standardization, and then apply on features.	CO 4
5	Summarize the properties of a normal density function, and justify why this density function is widespread.	Analyze	The learner will try to recall and understand the normal density function, and central limit theorem, and then justify.	CO 4
6	Discuss various methods of RE.package and justify how group() method is useful in searching for a phone number	Understand	The learner will try to recall the definitions of RegEx,understand the pythonRE.package and then explain group() method.	CO 4

7	Summarize the token based fuzzy matching methods and explain these methods are useful in removing duplicate words or set of words	Understand	The learner will try to recall fuzzy matching methods,understand the pyhton's FuzzyWuzzy library, and then summarize.	CO 4
8	Justify why inconsistencies in measurement scale, in data types and in data range must be corrected before data analysis task.	Analyze	The learner will try to recall possible data inconsistencies, understand their effect on data analysis, and then justify.	CO 4
9	Justify that mode might not be a good choice to replace the invaid data in a dataset.	Analyze	The learner will try to recall the definition of mode, and understand the normal density function, and central limit theorem, and then justify.	CO 4
10	Compose a python program for outlier detection using IQR method.	Apply	The learner will try to recall the definition of IQR, understand the pandas methods, and then develop a python code.	CO 4
PART-B LONG ANSWER QUESTIONS				
1	Compare and contrast the min-max and standard scalers for feature normalization.	Understand	The learner will try to recall and understand the feature scaling methods and compare their salient characteristics.	CO 4
2	Discuss about read_csv() method of pandas dataframe: application, input and output arguments.	Understand	The learner will try to recall and identify the file handling methods available for pandas dataframe, and then provide solution.	CO 4
3	Explain the steps in a typical wrangling with an example python code for each step.	Understand	The learner will try to recall the steps in wrangling and pandas, and then develope code for wrangling.	CO 4
4	Explain in detail, the possible causes of missing values, bad data, and duplicate values in raw data. Also review the possible solutions.	Apply	The learner will try to recall the natue of raw data, and identify several methods of data cleaning.	CO 4

5	Discuss about drop_duplicates() method of pandas dataframe: application, input and output arguments.	Apply	The learner will try to recall and identify the data cleanup methods available for pandas dataframe, and then provide solution.	CO 4
7	Discuss the role of quartiles in data wrangling with the help of an example. Give python code for computing quartiles using python built-in-methods or user-defined methods.	Apply	The learner will try to recall the definition of quartiles,understand the IQR, and then develop python code.	CO 4
8	Compose a python program for outlier detection using Std method. Give a sample python code for the same.	Apply	The learner will try to recall the definition of standard deviation (Std), understand its relation with cut-off value for outlier detection, and then develop python code.	CO 4
9	Compose a python program for computing Fuzzy full ratio and partial ratio scores. Also give the expected outputs for sample string inputs.	Apply	The learner will try to recall fuzzy ratios, understand the fuzzy matching, and then develop python code.	CO 4
10	Compose a python program for computing search and compute methods. Also give the expected outputs for sample string inputs.	Apply	The learner will try to recall Regular Expression concept and understand the RegEx matching, and then develop python code.	CO 4
11	Compose a python program for computing findall method. Also give the expected outputs for sample string inputs.	Apply	The learner will try to recall Regular Expression concept and understand the RegEx matching, and then develop python code.	CO 4
12	Explain DictReader module to deal with duplicate records.	Apply	The learner will try to recall Regular Expression concept and understand the DictReader module.	CO 4
13	Explain Normalizing and standardizing the data.	Apply	The learner will try to recall Normalizing and Standardizing the Data concept in detail.	CO 4

14	How to save the new clean data using SQLite into database.	Analyze	The learner will try to recall data storage process with SQLite database.	CO 4
15	Explain Zen of Python for documenting the code.	Apply	The learner will try to recall data scripting.	CO 4
16	How to create a File Command-Line Executable.	Apply	The learner will try to recall data scripting.	CO 4
17	Create a logical repository structure using README.md and .gitignore files to have an organized project folder	Analyze	The learner will try to recall data scripting.	CO 4
18	Compose a python program for importing dataset from UNICEF Website.	Analyze	The learner will try to recall data scripting.	CO 4
19	Differentiate unittest,nose and pytest tesing methodologies.	Apply	The learner will try to recall different python code testing frameworks.	CO 4
20	How to perform pytest testing on python code.	Apply	The learner will try to recall testing frameworks.	CO 4
PART-C SHORT ANSWER QUESTIONS				
1	Explain the pandas method useful for reading 2-dimensional data from an ascii file	Understand	The learner will try to recall and identify the file handling methods available in pandass, and then provide solution	CO 4
2	Discuss in brief the pandas method used for summarizing the staistics of a dataframe.	Understand	The learner will try to recall and identify the file handling methods available in pandass, and then provide solution	CO 4
3	What is the useful method to print data types contained in a dataframe. Explain in brief.	Understand	The learner will try to recall and identify the data summarization methods available in pandass, and then provide solution	CO 4
4	What are the methods used to extract the row and column labels of a pandas dataframe?	Remember	–	CO 4
5	Write a short note on method(s) used to treat the missed values of a pandas dataframe.	Remember	–	CO 4

6	Write a short note on contingency table of a dataframe.	Remember	–	CO 4
7	What is a fuzzy set? explain.	Remember	–	CO 4
8	Define the full and partial ratios used in fuzzy matching.	Remember	–	CO 4
9	Explain the RE module in python?	Remember	–	CO 4
10	How quartiles and percentiles are related? Explain in brief?	Remember	–	CO 4
11	Write a short note on finding duplicates?	Remember	–	CO 4
12	Differentiate Fuzz Matching and RegEx Matching?	Understand	–	CO 4
13	Define Normalizing of data with example?	Remember	–	CO 4
14	Write helper function to zip headers with rows of data?	Understand	–	CO 4
15	?	Remember	–	CO 4
16	Write a shore note on finding duplicates?	Remember	–	CO 4
17	Write a short note on finding missing data?	Remember	–	CO 4
18	Write steps for creation of README.md file?	Understand	–	CO 4
19	Write a short note on UNICEF dataset?	Remember	–	CO 4
20	Define Git version control and .gitigore files?	Understand	–	CO 4
MODULE IV				
DATA EXPLORATION AND ANALYSIS				
PART A-PROBLEM SOLVING AND CRITICAL THINKING QUESTIONS				
1	Discuss in detail the role of a cross tabling of attributes in data exploration. Give sample python code for the same	Apply	The learner will try to recall pandas methods,understand the concept of cross tabling, and then develop pyhton code to create a cross table.	CO 5

2	Justify that a box plot can be used for getting rank statistics of an attribute. Give sample python code for the same	Apply	The learner will try to recall boxplot and matplotlib library, understand the rank statistics and then then develop pyhton code to view box plot.	CO 5
3	Discuss in detail the role of a pivot table of attributes in a data exploration. Give python code for creating a pivot table.	Apply	The learner will try to recall pandas methods, understand the concept of pivot table, and then develop pyhton code to create the table.	CO 5
4	Develop a python code for estimating the joint, marginal and conditional probabilities of two attributes.	Apply	The learner will try to recall pandas methods,understand the definitions of joint, marginal and conditional probabilities, and then develop python code.	CO 5
5	Compare and contrast the pandas.merge() and DataFrame.merge() to merge multiple Dataframes.	Apply	The learner will try to recall pandas methods, understand the data merging process, and then compare both the pandas methods.	CO 5
6	Justify you ensure data integrity and quality during the import process?	Apply	The learner will try to recall and understand the concepts of Data Integrity and apply the Knowledge.	CO 5
7	Justify you handle missing or mismatched data when performing joins?	Apply	The learner will try to recall and understand the concepts of Joining numerous datasets and apply the knowledge to handle the missing data.	CO 5
8	Enumerate the benefits and potential challenges of using open-source platforms for publishing data?	Apply	The learner will try to recall and understand the concepts of Publishing the data and implement the knowledge.	CO 5
9	Enumerate some techniques for identifying and interpreting correlations between variables in a dataset?	Apply	The learner will try to recall and understand the concepts of Identifying Correlations and apply the knowledge.	CO 5

10	Explain the considerations you need to take into account when importing data from different file formats. Provide examples of potential challenges and how to address them.	Apply	The learner will try to recall and understand the methods of importing the data and applying the knowledge.	CO 5
PART-B LONG ANSWER QUESTIONS				
1	What is the importance of groupby() method of pandas in data exploration. Give sample python code for the same.	Apply	The learner will try to recall pandas methods, understand the attribute grouping, and then develop pyhton code to create attribute groups.	CO 5
2	Explain in detail, the Product-Moment Correlation Coefficient and its role in exploring the data attributes.	Apply	The learner will try to recall pandas methods, understand the Correlation, and then develop pyhton code to compute Correlation Coefficient between each pair of attributes	CO 5
3	Compare and contrast the contingency table and pivot table of attributes in a dataframe.	Understand	The learner will try to recall both contingency and table pivoting, understand pandas methods, and then compare their salience.	CO 5
4	Discuss in detail the role ofa pivot table of attributes in data exploration. Give sample python code for the same. The learner will try to recall table pivoting, understand pandas methods, and then develop python code.	Apply	The learner will try to recall table pivoting, understand pandas methods, and then develop python code.	CO 5
5	Discuss in detail the importance of Q-Q plots in data exploration. Give sample python code for the generating a Q-Q plot.	Apply	The learner will try to recall pandas methods, understand table pivoting, and then develop python code.	CO 5

6	Compare and contrast between covariance and correlation of data attributes.	Apply	The learner will try to recall pandas methods, understand covariance and correlation, and then compare both.	CO 5
7	Discuss the two methods of merging multiple dataframes. Give sample python code.	Apply	The learner will try to recall pandas methods, understand dataframe merging, and then develop python code.	CO 5
8	Distinguish between maximum absolute deviation (MAD) and Std methods of scaling of data attributes. Give sample python code for both the methods.	Apply	The learner will try to recall pandas methods, understand data scaling techniques, and then develop python code.	CO 5
9	Distinguish between the IQR and Std methods of outlier eilimination in a dataset. Give sample python code for both the methods	Apply	The learner will try to recall the definitions of quartiles and standard deviation(Std), understand data outliers, and then develop python code	CO 5
10	Create groupings with an example. Give sample python code for creating groupings	Apply	The learner will try to recall the groupings, creation of groupings and then develop python code	CO 5
11	Use agate library to filter data and give sample python ode for filtering data.	Apply	The learner will try to recall the separation of data and then develop python code	CO 5
12	Differentiate various presentations for data.	Apply	The learner will try to recall the data presentations	CO 5
13	Explain charting with matplotlib using sample python code.	Apply	The learner will try to recall the data visualization	CO 5
14	Explain charting with Bokeh library using sample python code.	Analyze	The learner will try to recall the data visualization	CO 5
15	Distinguish between matplotlib and Bokeh libraries using sample python code.	Analyze	The learner will try to recall the data visualization	CO 5

16	How to use pygal with agate table to show the worldwide child labor rates.	Understand	The learner will try to recall the data visualization	CO 5
17	Give a brief notes on presentation tools for data.	Understand	The learner will try to recall the presentation of data.	CO 5
18	Explain Ghost platform with detailed procedure.	Analyze	The learner will try to recall the data publication	CO 5
19	Explain GitHub pages version control system and Jekll framework.	Apply	The learner will try to recall the data visualization	CO 5
20	Explain Jupyter or IPython Notebook for python code sharing.	Analyze	The learner will try to recall the concept of code sharing	CO 5
PART-C SHORT ANSWER QUESTIONS				
1	What is time series data? Explain in brief.	Remember	–	CO 5
2	What are table functions useful in pandas?	Remember	–	CO 5
3	Explain the difference between independence and uncorrelatedness of attributes in a dataset.	Remember	–	CO 5
4	Write a short note on groupby() method of pandas.	Remember	–	CO 5
5	Explain the pearson's correlation coefficient in data wragling.	Remember	–	CO 5
6	Explain the crosstab() method of pandas.	Remember	–	CO 5
7	Explain why quartiles can be considered as rank statistics.	Remember	–	CO 5
8	Explain Exception Handling to communicate with errors.	Remember	–	CO 5
9	Write a short notes on agate library.	Remember	–	CO 5
10	Define JSON file structure with example.	Remember	–	CO 5
11	Explain avoiding storytelling pitfalls.	Remember	–	CO 5
12	Define data visualization using chart .	Remember	–	CO 5

13	Explain timeline and time series charts .	Remember	–	CO 5
14	Write a shore notes on Maps for data visualization.	Remember	–	CO 5
15	Define the usage of pygal library in data visualization.	Remember	–	CO 5
16	Write a short notes on Prezi presentation tool.	Remember	–	CO 5
17	Differentiate CSV and TSV file for presenting the data.	Remember	–	CO 5
18	List some open source platforms for staring new site up.	Remember	–	CO 5
19	Write a short notes on Ghost platform.	Remember	–	CO 5
20	Write a short notes on IPython Notebook.	Remember	–	CO 5
MODULE V				
WEB SCRAPING				
PART A-PROBLEM SOLVING AND CRITICAL THINKING QUESTIONS				
1	Justify that a quantile-quantile plot can be used for identifying gaussian attributes.	Analyse	The learner will try to recall the definition of quantiles and properties of gaussian variable, understand the quantile-quantile plot, and justify the statement.	CO 5
2	Analyse the box plot of a gaussian attribute, and bring out the salient features of such a plot.	Analyse	The learner will try to recall the characteristics of a box plot and properties of gaussian variable, analyse the plot and draw conclusions.	CO 6
3	Illustrate the key considerations when determining what information to scrape from a website. How do you decide on the scraping approach (e.g., specific elements or entire pages)?	Apply	The learner will try to recall and understand the concepts of what to scrape and how and apply the knowledge gained.	CO 6

4	Analyze the key considerations when determining what information to scrape from a website. How do you decide on the scraping approach (e.g., specific elements or entire pages)?	Apply	The learner will try to recall and understand the concepts of what to scrape and how and apply the knowledge gained.	CO 6
5	Illustrate the methods for retrieving web pages for scraping, including the use of libraries and HTTP requests.	Apply	The learner will try to recall and understand the concepts of Getting pages.	CO 6
6	Analyze the concept of a web spider and how Scrapy can be used to build one. What are the benefits of using a web spider?	Apply	The learner will try to recall and understand the Ghost.py library and screen reading using Ghost.py	CO 6
7	illastrate the process of crawling an entire website using Scrapy. What challenges may arise, and how can they be addressed?	Apply	The learner will try to recall and understand the process of Crawling whole websites with scrapy.	CO 6
8	Analyze Discuss the challenges and techniques for scraping data from web pages that heavily rely on JavaScript for content rendering.	Understand	The learner will try to recall the concept of Interaction with JavaScript.	CO 6
PART-B LONG ANSWER QUESTIONS				
1	Explain the role of stacked bar graph in data visualization. Give sample python code for creating a stacked bar graph.	Apply	The learner will try to recall the characteristics of a stacked bar graph,understand python graphics and develop code for drawing the graph.	CO 5
2	Explain how a scatter plot is useful in data visualization. Give sample python code for creating a scatter plot.	Apply	The learner will try to recall the characteristics of a scatter graph, understand python graphics and develop code for drawing the scatter plot.	CO 5

3	Explain how a histogram plot is useful in data visualization. Give sample python code for creating a scatter plot.	Apply	The learner will try to recall the procedure for buliding a histogram, understand python graphics and develop code for drawing the histogram plot.	CO 5
4	Explain how a multiline plot is useful in data visualization. Give sample python code for creating a scatter plot.	Apply	The learner will try to recall the procedure for buliding a multiline plot, understand python graphics and develop code for drawing the plot.	CO 5
5	Explain the Document Object Model (DOM) of a web page. What is its role in web scraping?	Understand	The learner will try to recall the definition of DOM, understand the web scraping principles and then explain the DOM role.	CO 6
6	Explain Web Crawling and Web Scraping. Compare and contrast the both.	Understand	The learner will try to recall and understand the concepts of Web Scraping and crawling, and then compare them.	CO 6
7	Explain the procedure of reading a Web Page with BeautifulSoup with the help of python code.	Understand	The learner will try to recall the web page layout, understand the features of BeautifulSoup library, and then develop python code.	CO 6
8	What you mean by "Spidering the Web"? Explain how to build web spider with Scrapy.	Apply	The learner will try to recall the Scrapy fraemwork, weg page layout, and then explain the solution.	CO 6
9	Explain how Selenium is used to read a Web Page using firefox browser using python. Also discuss briefly to extract text from bubbles.	Apply	The learner will try to recall the Selenium library, understand the features of the fire-fox browser and then explain the required procedure.	CO 6
10	Explain how to read a Web Page with LXML module in python.	Understand	The learner will try to recall the LXML module, understand the features of a web page and then explain the required procedure.	CO 6

11	What is browser-based parsing? What is the role selenium in this? Explain.	Understand	The learner will try to recall the concept of web browsing and parsing, understand the capabilities of selenium library, and then explain the answer.	CO 6
12	Explain page loading with a detailed steps.	Understand	The learner will try to recall the concept of web browsing and parsing, understand the capabilities of selenium library, and then explain the answer.	CO 6
13	Explain screen reading with Ghost.py WebKit.	Understand	The learner will try to recall the concept of web browsing and parsing, understand the capabilities of selenium library, and then explain the answer.	CO 6
14	How web spidering works with scrapy web spider tool.	Understand	The learner will try to recall the concept of web browsing and parsing, understand the capabilities of selenium library, and then explain the answer.	CO 6
15	How to use Scrapy shell to investigate content on the Scrapely page. Explain.	Understand	The learner will try to recall the concept of web browsing and parsing, understand the capabilities of selenium library, and then explain the answer.	CO 6
16	Justify, How the Internet Works and Why It's Breaking Your Script.	Understand	The learner will try to recall the concept of web browsing and parsing, understand the capabilities of selenium library, and then explain the answer.	CO 6
17	Distinguish between screen reading with selenium and screen reading with Ghost.py.	Understand	The learner will try to recall the concept of web browsing and parsing, understand the capabilities of selenium library, and then explain the answer.	CO 6

18	Explain DOM structure with detailed steps.	Understand	The learner will try to recall the concept of web browsing and parsing, understand the capabilities of selenium library, and then explain the answer.	CO 6
19	Explain scrapers creation in PhantomJS with GhostDriver.	Understand	The learner will try to recall the concept of web browsing and parsing, understand the capabilities of selenium library, and then explain the answer.	CO 6
20	Explain Spider Building with scrapy and scrapy class system.	Understand	The learner will try to recall the concept of web browsing and parsing, understand the capabilities of selenium library, and then explain the answer.	CO 6
PART-C SHORT ANSWER QUESTIONS				
1	What is BeautifulSoup in web scraping?	Remember	-	CO 6
2	What is the structure and markup of a web site?	Remember	-	CO 6
3	What is the find all() method of BeautifulSoup? explain.	Remember	-	CO 6
4	What is Ghost.py?	Remember	-	CO 6
5	What do you mean by "Spidering the Web"?	Remember	-	CO 6
6	What are different web spiders?	Remember	-	CO 6
7	Explain the function of a web crawler.	Remember	-	CO 6
8	Give the anatomy of a typical web page.	Remember	-	CO 6
9	What is a Cascading Style Sheet?	Remember	-	CO 6
10	What is Selenium?	Remember	-	CO 6
11	Write a short note on scree reading?	Remember	-	CO 6
12	Write a short notes Ghost.py webkit?	Remember	-	CO 6

13	Define Brower-based parsing?	Remember	-	CO 6
14	Write a short notes on crawling whole website with scrapy?	Remember	-	CO 6
15	Write python code to locate the input box?	Remember	-	CO 6
16	Write a short notes on Gost.py click method?	Remember	-	CO 6
17	Write python code for web page opening using selenium ?	Remember	-	CO 6
18	Explain python's pass usage?	Remember	-	CO 6
19	Write a short notes on Selenium browser class's implicitly$_{w}aitmethod$?	Remember	-	CO 6
20	Define ActionChains and execute$_{s}criptmethod$?	Remember	-	CO 6

Signature of Course Coordinator **HOD,CSE(DS)**
Ms J Sirisha, Assistant Professor