



SIDDHARTH INSTITUTE OF ENGINEERING & TECHNOLOGY
(AUTONOMOUS)

(Approved by AICTE, New Delhi & Affiliated to JNTUA, Ananthapuramu)

(Accredited by NBA for Civil, EEE, Mech., ECE & CSE)

Accredited by NAAC with 'A+' Grade)

Puttur -517583, Tirupathi District, A.P. (India)



QUESTION BANK (DESCRIPTIVE)

Subject with Code: Data Wrangling (23CS0935)

Course & Branch: CSM & CAI

Year & Sem: IV B.Tech & I Sem

Regulation: R23

UNIT I

Introduction to Data Wrangling and Data Acquisition

1	a)	Define Data Wrangling and state its main goal.	[L1,CO1]	[2M]
	b)	Give some examples of Semi-structured data formats.	[L1,CO1]	[2M]
	c)	Write the pandas code to read a file named students.xlsx.	[L1,CO1]	[2M]
	d)	Name the Python libraries used for Web Scraping.	[L1,CO1]	[2M]
	e)	What is the role of SQLAlchemy when working with pandas and databases?	[L1,CO1]	[2M]
2	a)	Explain the importance of the Data Wrangling.	[L2,CO1]	[5M]
	b)	Briefly Explain the use cases of Data Wrangling.	[L2,CO1]	[5M]
3		Describe the Data acquisition techniques using APIs and Web scraping used in data wrangling.	[L5,CO1]	[10M]
4	a)	Outline how to read data from CSV and Excel files using pandas.	[L4,CO1]	[5M]
	b)	Construct how to read data from JSON, and XML files using pandas Library.	[L4,CO1]	[5M]
5		Compare Structured, Semi-structured, and Unstructured data based on format and examples.	[L4,CO1]	[5M]
6	a)	Why is chunking required when working with large datasets and describe it in detail.	[L2,CO1]	[5M]
	b)	Illustrate how to connect to a database and execute a custom SQL query using SQLAlchemy and pandas.	[L4,CO1]	[5M]
7		Demonstrate the use of pandas and requests for data acquisition.	[L5,CO1]	[10M]
8		Explain the steps to scrape data from a webpage using BeautifulSoup.	[L3,CO1]	[10M]
9		Investigate the steps of Exploratory Data Analysis before cleaning.	[L6,CO1]	[10M]
10		Examine why loading large datasets directly causes memory errors.	[L5,CO1]	[10M]

UNIT II

Handling Missing, Noisy and Inconsistent data

1	a)	List the techniques used for imputing missing values.	[L1,CO2]	[2M]
	b)	Mention the methods used for detecting outliers.	[L1,CO2]	[2M]
	c)	Define duplicate data. Why should duplicate records be removed?	[L1,CO2]	[2M]
	d)	What is normalization?	[L1,CO2]	[2M]
	e)	Which plots are commonly used to detect outliers?	[L1,CO2]	[2M]
2		Explain different types of missing data. Discuss various techniques used for imputing missing values with suitable examples.	[L2,CO2]	[10M]
3		Discuss different methods for handling inconsistent data, including date formats, text inconsistencies, and unit conversions. Illustrate with suitable examples.	[L,CO2]	[10M]
4	a)	Compare and evaluate deletion methods and imputation methods for handling missing values.	[L5,CO2]	[5M]
	b)	Investigate the causes of missing values in real-world datasets.	[L5,CO2]	[5M]
5		Analyze the impact of duplicate and irrelevant data on data quality and propose an effective strategy to eliminate them before analysis.	[L6,CO2]	[10M]
6		Examine different outlier detection techniques and determine which method is more suitable for different types of datasets.	[L5,CO2]	[10M]
7	a)	Assess various methods for treating outliers.	[L5,CO2]	[5M]
	b)	Discuss about normalization techniques.	[L4,CO2]	[5M]
8	a)	Design a solution for standardization numeric features before training.	[L6,CO2]	[5M]
	b)	Compare and Contrast Normalization and standardization techniques.	[L6,CO2]	[5M]
9		Discuss the role of Regular Expressions in Text Cleaning.	[L4,CO2]	[10M]
10		Interpret different visualization techniques used for identifying missing values and outliers, and infer how these visualizations support effective data preprocessing.	[L2,CO2]	[10M]

UNIT III

Data Transformation and Feature Engineering

1	a)	What is binning? Mention two advantages of discretization.	[L1,CO3]	[2M]
	b)	Why is feature engineering considered an important step in machine learning?	[L1,CO3]	[2M]
	c)	Define a derived feature with a suitable example.	[L1,CO3]	[2M]
	d)	List any four features that can be extracted from date and time data.	[L1,CO4]	[2M]
	e)	Differentiate between One-Hot Encoding and Label Encoding.	[L1,CO4]	[2M]
2		Identify the appropriate data type conversions required for a given dataset and outline the parsing steps needed to prepare it for analysis.	[L3,CO3]	[5M]
3	a)	Validate the effectiveness of selected features using appropriate feature extraction techniques.	[L3,CO3]	[5M]
	b)	Categorize the feature extraction from text, dates and strings.	[L4,CO3]	[5M]
4		Explain about the One-Hot Encoding and Label Encoding, and select the appropriate technique for different categorical variables.	[L5,CO3]	[10M]
5		Organize continuous data into meaningful intervals using binning and discretization techniques, and demonstrate their advantages.	[L6,CO3]	[10M]
6		Inspect the impact of different discretization methods on data quality and model performance.	[L4,CO3]	[5M]
7	a)	Examine the role of data aggregation in the business intelligence.	[L4,CO3]	[5M]
	b)	Summarize about the data grouping in the decision-making applications.		[5M]
8	a)	Compare and contrast the Data Aggregation and Grouping.		
	b)	How will you handle imbalanced data in the feature engineering.	[L3,CO3]	[5M]
9		Judge the effectiveness of Pivoting, Melting and reshaping data in addressing class imbalance using appropriate scenarios.	[L5,CO3]	[10M]
10		Design a complete feature engineering workflow for creating derived features and feature selection.	[L5,CO4]	[10M]

UNIT IV
Data Integration, Joining, and Workflows

1	a)	Identify the purpose of concatenating DataFrames.	[L1,CO5]	[2M]
	b)	State the meaning of referential integrity.	[L1,CO5]	[2M]
	c)	Recall any two causes of schema mismatches.	[L1,CO5]	[2M]
	d)	Classify the major stages of an end-to-end data wrangling pipeline.	[L4,CO5]	[2M]
	e)	Interpret the importance of documentation in data integration.	[L1,CO5]	[2M]
2		Demonstrate how to merge and join datasets using Inner, Left, Right, and Outer Joins.	[L4,CO5]	[10M]
3		Illustrate the process of concatenating and appending DataFrames with suitable examples.	[L3,CO5]	[10M]
4	a)	Explain how the data consistency improves data integration	[L2,CO5]	[5M]
	b)	Examine the impact of referential integrity violations on database quality.	[L3,CO5]	[5M]
5		Use appropriate methods to resolve schema mismatches while integrating datasets from multiple sources.	[L6,CO5]	[10M]
6	a)	Compare Inner Join, Left Join, Right Join, and Outer Join based on output and applications.	[L4,CO5]	[5M]
	b)	Critique the advantages and limitations of reusable data wrangling functions.	[L5,CO5]	[5M]
7		Generate designing of the Reusable data wrangling functions in data joining and workflows.	[L3,CO5]	[5M]
8	a)	Design an automated workflows with functions and pipelines.	[L6,CO5]	[5M]
	b)	Define data lineage and why its important in data integration.	[L4,CO5]	[5M]
9		Justify the need for documentation in collaborative data engineering projects.	[L5,CO5]	[10M]
10		Describe the case study of End-to-End data wrangling pipeline.	[L5,CO5]	[10M]

UNIT V
Tools, Libraries, and Case Studies in Data Wrangling

1	a)	Point out two common pitfalls encountered during data wrangling.	[L1,CO5]	[2M]
	b)	Suggest two evaluation metrics for assessing a cleaned dataset before machine learning.	[L1,CO5]	[2M]
	c)	State two advantages of using OpenRefine for data cleaning.	[L1,CO6]	[2M]
	d)	List any four advanced features of Pandas used in data wrangling.	[L1,CO5]	[2M]
	e)	Recall the importance of versioning in reproducible data pipelines.	[L1,CO5]	[2M]
2		Explain the advanced data manipulation techniques available in Pandas and NumPy with suitable examples.	[L6,CO6]	[10M]
3		Analyze the advantages and limitations of Pandas, NumPy, Pyjanitor, Dask, and Polars in large-scale data wrangling.	[L6,CO6]	[10M]
4	a)	Describe how advanced Pandas operations can be used to clean and transform a large dataset.	[L4,CO5]	[5M]
	b)	Discuss the workflow of OpenRefine for detecting and correcting inconsistent data values.	[L5,CO6]	[5M]
5		Demonstrate the use of Pyjanitor functions to simplify data cleaning and preprocessing with appropriate examples.	[L3,CO5]	[5M]
6	a)	Justify the selection of SQL or NoSQL databases for different real-world data wrangling scenarios.	[L5,CO5]	[5M]
	b)	Illustrate how SQL queries can be integrated into a data wrangling pipeline for preprocessing data.	[L4,CO5]	[5M]
7		Analyze the common pitfalls in data wrangling and propose suitable strategies to overcome them.	[L6,CO6]	[10M]
8		Examine a healthcare data wrangling case study and discuss the preprocessing steps required before predictive analysis.	[L4,CO5]	[10M]
9		Estimate the role of reproducibility and versioning in data pipelines.	[L3,CO5]	[10M]
10		Develop a complete workflow for building a clean dataset for machine learning, including data collection, cleaning, transformation, validation, and evaluation.	[L6,CO6]	[10M]