UNIT-V

10	a)	Demonstrate how heatmaps and scatter plots can be used to extract insights from a dataset.	L4	CO4	7 M
	b)	Discuss the limitations of these visualization techniques.	L4	CO4	3 M
		OR			
11	a)	Evaluate the effectiveness of different data visualization techniques in understanding complex datasets.	L4	CO4	6 M
	b)	Discuss how these techniques facilitate communication between data scientists and stakeholders.	L4	CO4	4 M

Code: 23ES1306

PVP 23

II B.Tech - I Semester – Regular Examinations - DECEMBER 2024

INTRODUCTION TO DATA SCIENCE
(DATA SCIENCE)

Duration: 3 hours

Max. Marks: 70

Note: 1. This question paper contains two Parts A and B.
2. Part-A contains 10 short answer questions. Each Question carries 2 Marks.
3. Part-B contains 5 essay questions with an internal choice from each unit. Each Question carries 10 marks.
4. All parts of Question paper must be answered in one place.

BL – Blooms Level

CO – Course Outcome

PART – A

		BL	CO
1. a)	Describe the main tasks involved in Data Science.	L2	CO1
b)	Explain the importance of business understanding in the Data Science lifecycle.	L2	CO1
c)	Identify the types of data formats commonly used in Data Science.	L2	CO1
d)	Discuss the role of APIs in data acquisition.	L2	CO2
e)	Explain the process of data cleaning with an example.	L2	CO1
f)	Compare smoothing and Normalization.	L2	CO1
g)	Demonstrate about interquartile range and its significance.	L2	CO1
h)	Discuss various distribution shapes.	L2	CO1
i)	Discuss any two data visualization techniques to summarize data.	L2	CO1
j)	Demonstrate the use of box plots in data analysis.	L2	CO1

PART – B

			BL	CO	Max. Marks
		UNIT-I			
2	a)	Explain the major tasks involved in the Data Science.	L2	CO1	6 M
	b)	Explain the CRISP-DM methodology.	L2	CO1	4 M
		OR			
3	a)	Explain in detail about the role of data preparation in a Data Science project.	L2	CO1	4 M
	b)	Discuss about the applications of Data Science.	L2	CO1	6 M
		UNIT-II			
4	a)	Analyze the challenges of acquiring data through APIs and web scraping.	L4	CO4	6 M
	b)	Discuss about different data collection methods.	L2	CO2	4 M
		OR			
5	a)	Compare structured, semi-structured, and unstructured data.	L3	CO2	5 M
	b)	Discuss the advantages and challenges of working with each type.	L2	CO2	5 M
		UNIT-III			
6	a)	A dataset contains missing values as follows: [45, 50, NaN, 55, NaN, 65]. Apply atleast four different appropriate data cleaning techniques and compare them.	L4	CO4	8 M
	b)	Justify your choice of technique and its impact on the data analysis process.	L4	CO4	2 M
		OR			
7	a)	A dataset contains the following attribute values, [45 50 55 60 65 70 75 80 85 90]. Perform data transformation by applying normalization using min-max scaling and z-score scaling.	L3	CO3	8 M
	b)	Compare the results and discuss their impact on data analysis.	L3	CO3	2 M
		UNIT-IV			
8	a)	Central tendency and variability are important measures in analysis, discuss how these measures aids in decision-making.	L3	CO3	4 M
	b)	Given the dataset [1, 2, 4, 7, 8, 10, 10], calculate the standard deviation and variance. Discuss their significance in understanding data variability.	L3	CO3	6 M
		OR			
9	a)	Given the following two sets of data: X = [12, 15, 14, 10, 18] and Y = [22, 25, 21, 20, 23], calculate Spearman's rank correlation coefficient and interpret the result.	L3	CO3	6 M
	b)	Interpret what the correlation coefficients in general imply about the relationships between variables.	L3	CO3	4 M

Code: 20ES1306

PVP 23

II B.Tech. – I Sem- Regular Examinations
DECEMBER 2024
CSE (Data Science)
INTRODUCTION TO DATA SCIENCE

1. a) Tasks in Data Science 2M

b) Importance of Business Understanding 2M

c) Common data formats 2M

d) Role of API 2M

e) process of data cleaning 2M

f) Comparison 2M

g) Definition of IQR 2M

h) any two shapes 2M

i) any two techniques 2M

j) any two uses 2M

2 a) Explanation 6M

b) Explanation 4M.

3 a) Explanation 4M.

b) Any six applications 6M

4 a) Explanation 6M

b) Explanation 4M.

5 a) Comparison 5M

b) Explanation 5M.

6 a) Any four Data cleaning Techniques 8M

b) Justification 2M.

7 a) Min-max & Z- score scaling 8M.

b) Comparison 2M

8 a) Explanation 4M

b) Calculation of Variance & Standard deviation 6M.

9 a) Calculation of Spear man's correlation coefficient 6M

b) Interpretation 4M.

10 a) Explanation 7M

b) Explanation 3M.

11 a) Explanation 6M

b) Explanation 4M.

1. a) Describe the main tasks involved in Data Science. [L2:CO1] 2M.

- **Tasks in Data Science:**
  1. Data Collection
  2. Storing data
  3. Data Processing
  4. Exploratory Data Analysis
  5. Data Modeling

1. b) Explain the importance of business understanding in the Data Science lifecycle. [L2:CO1] 2M.

- **Business understanding** is a critical first step in the data science lifecycle because it ensures that the data science efforts align with the business goals and address real-world problems effectively. Without a clear understanding of the business context, the model may focus on irrelevant metrics or solve the wrong problem.

1. c) Identify the types of data formats commonly used in Data Science. [L2:CO1] 2M.

- **Common data formats:**
  1. CSV
  2. JSON
  3. XML

1. d) Discuss the role of APIs in data acquisition. [L2:CO1] 2M.

- APIs (Application Programming Interfaces) facilitate data acquisition by providing a structured and standardized way to access, retrieve, and interact with data from different sources.

1. e) Explain the process of data cleaning with an example. [L2:CO1] 2M.

- **Data cleaning:**
  1. Handle missing values.
  2. Data Smoothing Techniques.
- Example: Replacing missing income values in a dataset with the average income.

1. f) Compare smoothing and normalization. [L2:CO1] 2M.

- **Smoothing:** Reducing noise in the dataset.
- **Normalization:** Scaling data to a specific range (e.g., [0,1]).

1. g) Demonstrate about interquartile range and its significance. [L2:CO1] 2M.

- **Interquartile Range (IQR):**

- IQR = Q3 - Q1 (difference between the 75th and 25th percentiles).
- Used to detect outliers

1. h) Discuss various distribution shapes. **[L2:CO1] 2M.**

2 marks for any two shapes below

1. Positively Skewed (Right-skewed)
2. Negatively Skewed (Left-skewed)
3. Symmetrical, bell-shaped curve.

1. i) Discuss any two data visualization techniques to summarize data. **[L2:CO1] 2M.**

2 marks for any two techniques below

1. Histograms
2. Box Plot
3. Bar Plot/ Bar chart
4. Violin Plot

1. j) Demonstrate the use of box plots in data analysis. **[L2:CO1] 2M.**

- Box Plot visualizes data distribution using five-number summary (min, Q1, median, Q3, max).
- Identifies outliers.

PART-B

UNIT-I

2 a) Explain the major tasks involved in Data Science. **[L2:CO1] 6M.**

- The major tasks involved in Data Science :

1. **Data Collection:**
   - Gathering raw data from various sources (databases, APIs, web scraping, sensors).
   - Tools: SQL, APIs, Python libraries like requests and BeautifulSoup.
2. **Storing Data:**
   - Store the collected data in a way that it can be easily accessed and used for analysis.
   - Steps: Choose Storage Solutions, Data Organization, Data Security, Backup and Recovery
3. **Data Processing:**
   - Handling missing values, removing duplicates, and correcting data types.
   - Tools: Pandas library (Python), NumPy, data wrangling techniques.
4. **Exploratory Data Analysis (EDA):**
   - Analyzing data to identify patterns, trends, and outliers in datasets.
   - Tools: Matplotlib, Seaborn, and statistical techniques.
5. **Data Modeling:**
   - Applying machine learning algorithms to build predictive models.

- Steps: selection of the model and variables, execution of the model, and performing diagnostics and comparisons to improve model performance.

2 b) Explain the CRISP-DM methodology. [L2:CO1] 4M.

- **CRISP-DM (Cross Industry Standard Process for Data Mining)** is a framework for data science projects with 6 phases:

1. **Business Understanding:**
   - The first step in the data science life cycle is understanding the business problem you are trying to solve.
2. **Data Understanding:**
   - Collect and explore the data to solve the business problem.
3. **Data Preparation:**
   - Data preparation is about cleaning and transforming the raw data into a format suitable for analysis.
4. **Modeling:**
   - In this stage, machine learning or statistical models are built and trained to learn patterns from the data.
5. **Evaluation:**
   - After building the model, it is essential to evaluate how well it performs before deploying it into production.
6. **Deployment:**
   - Once the model has been evaluated and is performing well, it is deployed into a production environment where it can be used for making real-world decisions.

3 a) Explain in detail about the role of data preparation in a Data Science project. [L2:CO1] 4M.

- **Role of Data Preparation:**
  - Data preparation is about cleaning and transforming the raw data into a format suitable for analysis.

  Tasks:

  - Handle missing values.
  - Correct errors and inconsistencies in the data.
  - Normalize or standardize the data where needed.

3 b) Discuss about the applications of Data Science. [L2:CO1] 6M.

6 marks for any six applications below

1. **Finance and Banking :** Fraud detection, risk analysis, and stock price prediction, Customer Segmentation and Personalization etc.
2. **Marketing and Customer Behavior Analysis:** Customer segmentation, Predictive Analytics, Churn prediction etc.
3. **Transportation and Logistics:** Route optimization, Predictive Maintenance, Demand forecasting etc.
4. **Healthcare & Medicine:** Predicting diseases, personalized treatments, etc.
5. **Social media:** Sentiment analysis, User behavior analysis etc.
6. **E-commerce:** Recommendation systems etc.
7. **Cyber Security:** Anomaly detection, Intrusion detection etc.
8. **Agriculture:** Soil Health Assessment, Crop Disease and Pest Detection etc.
9. **Manufacturing:** Predictive maintenance, Quality control & defect detection etc.

UNIT-II

4 a) Analyze the challenges of acquiring data through APIs and web scraping. [L4:CO4] 6M.

API

APIs are mechanisms that enable two software components to communicate with each other using a set of definitions and protocols.

APIs provide a structured way to request and receive data from various sources, such as web services, databases, and applications. APIs allow developers to access specific data without having to deal with the underlying implementation details.

Challenges:

- Rate limits: Limited API calls per day.
- Authentication: Need API keys or tokens.
- Data inconsistency: API changes may break workflows.
- Cost: Paid APIs may have limited access in free tiers.

Web Scraping

Scraping is the process of extracting data from websites and converting it into a structured format, such as CSV, JSON, or XML. This technique is particularly useful for gathering unstructured data from the internet and organizing it for further analysis.

Challenges:

- Websites may have anti-scraping mechanisms like CAPTCHAs or IP blocking.
- Legal and ethical concerns, as some websites prohibit scraping in their terms of service.
- Handling dynamic websites that load content with JavaScript.

4 b) Discuss about different data collection methods. [L2:CO1] 4M.

1. **Surveys**: Collect structured data directly from individuals through questionnaires or forms.
2. **Experiments**: Gather data by conducting controlled tests to study cause-and-effect relationships.
3. **Sensor Data**: Obtain real-time measurements from devices like IoT sensors, wearables, or environmental monitors.
4. **Social Media Data**: Extract user-generated content and activity metrics from platforms like Twitter, Instagram, and Facebook.
5. **Transactional Data**: Capture data from business operations such as purchases, payments, and customer interactions.

5 a) Compare structured, semi-structured, and unstructured data. [L3:CO2] 5M.

1. Structured Data

Definition: Structured data is highly organized and easily searchable.

Characteristics:

- **Schema:** Defined and consistent (e.g., tables with rows and columns).

- **Format:** Typically numerical or categorical data.

2. Semi-Structured Data

Definition: Semi-structured data does not conform to a rigid schema but still has some organizational properties that make it easier to analyze than unstructured data.

Characteristics:

- **Schema:** Flexible and partially defined.
- **Format:** Often text-based with metadata or markers.

3. Unstructured Data

Definition: Unstructured data lacks a predefined format or structure, making it more challenging to collect, process, and analyze.

Characteristics:

- **Schema:** None; highly variable and free-form.
- **Format:** Text, images, videos, audio, etc.

5 b) Discuss the advantages and challenges of working with each type. [L2:CO2] 5M.

1. **Structured Data:**
   - Structured data is highly organized and easily searchable.
   - Advantage: Easy to analyze, query, and store.
   - Challenge: Limited flexibility for complex data.
2. **Semi-structured Data:**
   - Semi-structured data does not conform to a rigid schema but still has some organizational properties that make it easier to analyze than unstructured data.
   - Advantage: Flexible, supports hierarchical data.
   - Challenge: Parsing can be complex.
3. **Unstructured Data:**
   - Unstructured data lacks a predefined format or structure, making it more challenging to collect, process, and analyze.
   - Advantage: Rich insights from diverse data types.
   - Challenge: Complex to process and requires advanced techniques like NLP.

UNIT-III

6 a) A dataset contains missing values as follows: [45, 50, NaN, 55, NaN, 65]. Apply at least four different appropriate data cleaning techniques and compare them. [L4:CO4] 8M.

8 marks for any four techniques below

1. **Dropping Columns:**
   - Remove entire columns with many missing values.
   - Result: [45, 50, 55, 65].
2. **Imputation with Mean:**
   - Replace NaN with the mean value: Mean = (45+50+55+65)/4 = 53.75.
   - Result: [45, 50, 53.75, 55, 53.75, 65].
3. **Imputation with Median:**
   - Replace NaN with the median: Median = (45, 50, 55, 65) → 52.5.
   - Result: [45, 50, 52.5, 55, 52.5, 65].

4. **Imputing with Constant Values**
   - Fill with Zero: Common for numeric data.
   - Fill with "Unknown": Common for categorical data.
5. **Forward Fill:**
   - Use the last known value to fill NaNs.
   - Result: [45, 50, 50, 55, 55, 65].
6. **Backward Fill**:
   - Use the next known value to fill NaNs.
   - Result: [45, 50, 55, 55, 65, 65].
7. **Interpolation:**
   Estimate values based on neighboring values.

Comparison:

- Dropping Columns results in loss of information.
- Mean/median imputation is useful for numerical data.
- Forward fill works well for sequential data but may introduce bias.

6 b) Justify your choice of technique and its impact on the data analysis process. [L4:CO4] 2M.

Chosen Technique: Median Imputation

1. Handles Skewness: Median is robust to outliers and skewed data, ensuring a more accurate representation of the dataset.
2. Preserves Dataset Size: Unlike deletion, no data is lost, which is crucial for small datasets.

7 a) Perform data transformation on the given dataset [45, 50, 55, 60, 65, 70, 75, 80, 85, 90] using Min-Max Scaling and Z-Score Scaling. [L3:CO3] 8M.

1. Min-Max Scaling:

The formula is:

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A.$$

Min-max normalization.

Steps:

1. From given dataset, min=45 and max=90.
2. Apply the formula for each value.

Result after Min-Max Scaling:
[0, 0.11, 0.22, 0.33, 0.44, 0.56, 0.67, 0.78, 0.89, 1]

2. Z-Score Scaling:

The formula is:

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A}$$

(OR)

$$z = \frac{x - \mu}{\sigma}$$

Steps:

1. Calculate the mean μ=sum of all values\number of values.
2. Calculate the standard deviation σ using:

$$\sigma = \sqrt{\frac{1}{N}\sum(x_i - \mu)^2}$$

3. Apply the Z-Score formula for each value.

Result after Z-Score Scaling:
[-1.53, -1.19, -0.85, -0.51, -0.17, 0.17, 0.51, 0.85, 1.19, 1.53]

7 b) Compare the results and discuss their impact on data analysis. [L3:CO3] 2M.

- Min-Max Scaling: Rescales the data between 0 and 1. It is sensitive to outliers.
- Z-Score Scaling: Standardizes data around a mean of 0 and standard deviation of 1. It is robust to outliers.

8 a) Discuss central tendency and variability as important measures in analysis and how they aid in decision-making. [L3:CO3] 4M.

1. **Central Tendency:**
   - Describes the center of the data. Key measures:
     - **Mean (Average):** Sum of values divided by count.
     - **Median:** Middle value of the data when sorted.
     - **Mode:** Most frequent value.
2. **Variability (Spread):**
   - Describes the dispersion of data. Key measures:
     - **Range:** Difference between max and min.
     - **Variance and Standard Deviation:** Measure how far data points deviate from the mean.
     - **Interquartile Range (IQR):** Spread of the middle 50% of data.

8 b) Given the dataset [1, 2, 4, 7, 8, 10, 10], calculate the standard deviation and variance. [L3:CO3] 6M.

Step 1: Calculate the Mean (μ):

μ=Sum of values/No. of values

Mean (μ)=(1+2+4+7+8+10+10)/7

=42/7

=6

Step 2: Calculate Variance (σ2):
The formula is:

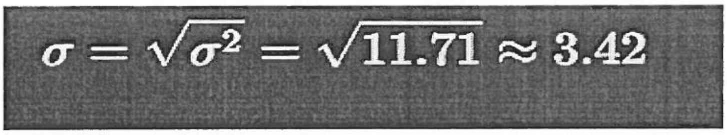
$$\hat{\sigma}^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1}$$

The population formula divides by N, while the sample formula divides by n-1

Steps:

1. Calculate the squared differences
2. Sum the squared differences: 25+16+4+1+4+16+16=82
3. **Divide by N (number of values): σ2=82/7≈11.71**
   (or)
   Divide by n-1: σ2=82/6≈13.6

Step 3: Calculate Standard Deviation (σ):



$$\sigma = \sqrt{\sigma^2} = \sqrt{11.71} \approx 3.42$$

(OR)

$$\sigma = \sqrt{13.6} \approx 3.69$$

Result:

- Variance = 11.71 (or) 3.42
- Standard Deviation = 13.6 (or) 3.69

9 a) Calculate Spearman's rank correlation coefficient for X = [12, 15, 14, 10, 18] and Y = [22, 25, 21, 20, 23]. [L3:CO3] 6M.

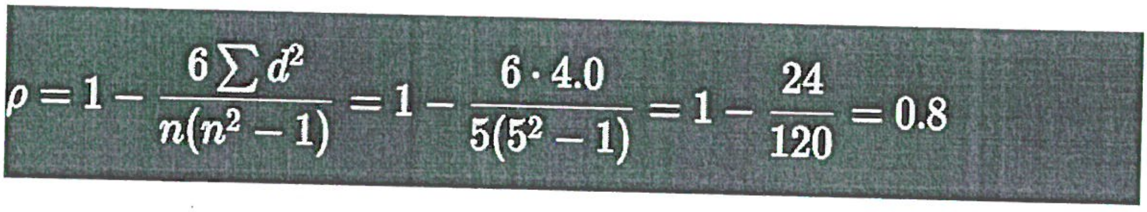
1. **Rank the values of X and Y & Find d_i^2**

X	Rank (X)	Y	Rank (Y)	d_i	d_i^2
12	2	22	3	-1	1
15	4	25	5	-1	1

14	3	21	2	1	1
10	1	20	1	0	0
18	5	23	4	1	1

d$_i$=Rank(X)−Rank(Y)

2. **Apply Spearman's Formula:**



$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \cdot 4.0}{5(5^2-1)} = 1 - \frac{24}{120} = 0.8$$

Final Answer: The Spearman's rank correlation coefficient is **0.8**.

Interpretation: There is a strong positive correlation between X and Y.

9 b) Interpret what the correlation coefficients imply about relationships. [L3:CO3] 4M.

- Positive Correlation (ρ > 0)
- Negative Correlation (ρ < 0)
- Zero Correlation (ρ ≈ 0)
- Here The Spearman's rank correlation coefficient is **0.8**.
- Interpretation: There is a strong positive correlation between X and Y.

10 a) Demonstrate how heatmaps and scatter plots can be used to extract insights from a dataset. [L3:CO3] 7M.

- **Heatmaps:**
  - Heatmaps represent data using colour intensity. They are ideal for showing correlations and patterns in large datasets.
  - Key Elements of a Heatmap: Grid Cells, Colour Gradient, X-axis and Y-axis
  - Grid Cells: The heatmap consists of a grid, where each cell represents a unique pair of x and y values (e.g., rows and columns).
  - Colour Gradient: Each cell is coloured based on its corresponding data value, with a colour gradient indicating the range of values.
  - X-axis and Y-axis: The axes represent the variables being compared.
  - For example, a heatmap can reveal how strongly different features of a dataset correlate with each other. Darker or lighter areas indicate stronger or weaker relationships.
  - Heatmaps are used for exploratory data analysis, particularly for understanding large matrices, e.g., correlation matrices.
- **Scatter Plots:**
  - Scatter plots display values for two variables as points on a 2D graph.
  - Key Elements of a Scatter Plot: Data Points, X-axis, Y-axis
  - Data Points: Each point on the scatter plot represents a single observation in the dataset. The coordinates of the point are defined by the values of the two variables being plotted.
  - X-axis: Represents the independent variable (or explanatory variable).
  - Y-axis: Represents the dependent variable (or response variable).

- By examining the distribution and patterns of the points, one can detect trends (e.g., positive/negative relationships), outliers, and clusters.
- They are used to understand relationships between variables, such as predicting outcomes (e.g., height vs. weight).

10 b) Discuss the limitations of these visualization techniques. [L3:CO3] 3M.

- **Heatmaps**:
  - Limited to summarizing data and showing patterns; exact values are hard to interpret.
  - Not suitable for small datasets.
  - Correlation does not imply causation.
- **Scatter Plots**:
  - Hard to interpret with large datasets due to overlapping points.
  - Only effective for 2D relationships.
  - Does not convey causation, only association.

11 a) Evaluate the effectiveness of different data visualization techniques in understanding complex datasets. [L4:CO4] 6M.

Plot Type	**Best For**	**Displays**	**Drawbacks**
Histogram	Distribution of a single variable	Frequency	Sensitive to bin size
Box Plot	Summarizing distribution & outliers	Five-number summary	Lacks distribution detail
Bar Plot	Comparing categorical data	Counts/proportions	Not suitable for continuous data
Scatter Plot	Relationships between variables	Correlations	Hard to use with dense data
Line Plot	Time series or trends	Sequential changes	Poor for non-sequential data
Heat Map	Dense correlations	Intensity/matrix	Overwhelming with many values
Pair Plot	Relationships in multivariate data	Pairwise scatter + histograms	Overlap in large datasets
Violin Plot	Distribution + comparison	Shape + summary	Complex for some viewers

These techniques help break down large datasets into digestible insights, improving understanding and decision-making.

11 b) Discuss how these techniques facilitate communication between data scientists and stakeholders. [L4:CO4] 4M.

- **Clarity**: Visualizations simplify complex data into easy-to-understand charts and graphs for non-technical audiences.
- **Improved Decision-Making**: By presenting key findings visually, stakeholders can make quicker and more informed decisions.
- **Storytelling**: Visualization techniques help communicate trends, patterns, and insights effectively.
- **Focus on Insights**: Instead of raw data, visuals highlight key takeaways that matter most to stakeholder.