U.S.N.										

B.M.S. College of Engineering, Bengaluru-560019

Autonomous Institute Affiliated to VTU

June 2025 Semester End Main Examinations

Programme: B.E. **Semester: V**

Branch: Computer Science and Engineering **Duration: 3 hrs.**

Course Code: 23CS5PCDEV **Max Marks: 100**

Course: Data Exploration and Visualization.

Instructions: 1. Answer any FIVE full questions, choosing one full question from each unit.
2. Missing data, if any, may be suitably assumed.

Important Note: Completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages. Revealing of identification, appeal to evaluator will be treated as malpractice.

<table>
<tr><th></th><th></th><th>UNIT - I</th><th>CO</th><th>PO</th><th>Marks</th></tr>
<tr><td>1</td><td>a)</td><td>List out the different measurement scales in data exploratory analysis explaining each of them with an example.</td><td>CO1</td><td>PO1</td><td>10</td></tr>
<tr><td></td><td>b)</td><td>Discuss the various steps performed in EDA and explain each process of the EDA in detail?</td><td>CO1</td><td>PO1</td><td>10</td></tr>
<tr><td></td><td></td><td><b>OR</b></td><td></td><td></td><td></td></tr>
<tr><td>2</td><td>a)</td><td>Describe the aims of Exploratory data analysis and differentiate between exploratory and confirmatory data analysis.</td><td>CO1</td><td>PO1</td><td>10</td></tr>
<tr><td></td><td>b)</td><td>With a neat diagram explain the classification of Exploratory Data Analysis.</td><td>CO1</td><td>PO1</td><td>10</td></tr>
<tr><td></td><td></td><td><b>UNIT - II</b></td><td></td><td></td><td></td></tr>
<tr><td>3</td><td>a)</td><td>Perform the following Transformation operations by considering two pandas data frames as shown below and write the output as well as operation code.
a) Inner Join b) Left Outer Join c) Full Outer join d) Right Outer Join e) append
<table><tr><th>ID</th><th>NAME</th></tr><tr><td>1</td><td>Alice</td></tr><tr><td>2</td><td>Bob</td></tr><tr><td>3</td><td>Charlie</td></tr></table>
<table><tr><th>ID</th><th>Age</th></tr><tr><td>2</td><td>25</td></tr><tr><td>3</td><td>30</td></tr><tr><td>4</td><td>28</td></tr></table>
</td><td>CO2</td><td>PO2</td><td>10</td></tr>
<tr><td></td><td>b)</td><td>Illustrate any five Transformation techniques applied in Data Transformation with an example for each technique.</td><td>CO1</td><td>PO1</td><td>10</td></tr>
<tr><td></td><td></td><td><b>OR</b></td><td></td><td></td><td></td></tr>
<tr><td>4</td><td>a)</td><td>Apply the concept of Discretization for data “height” by creating four bins and apply binning technique for data ages for following data shown</td><td>CO2</td><td>PO2</td><td>8</td></tr>
</table>

below. Write the Python code to create bins of equal width and equal frequency distribution.

height = [10, 20,31,54,51,15, 18, 34, 41,53]

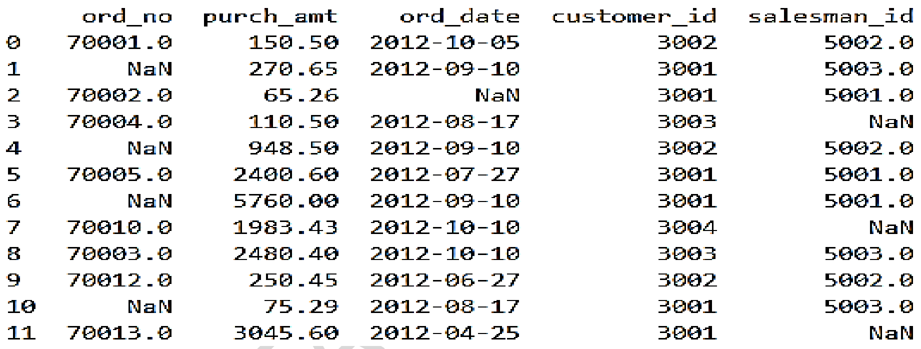
b) Suppose you have a dataset containing information about customers' purchases at a store. The dataset (customer_data.csv) includes columns: 'Customer_ID', 'Age', 'Gender', 'Purchase_Amount'. Your task is to perform random sampling to select a subset of 20 customers from this dataset for a survey. *CO2* *PO2* **6**

c) Apply the concept Binning technique for data ages for following data shown below. Write the Python code to create bins categorize them into different age groups ages = [22, 35, 47, 50, 28, 19, 65, 37, 42, 51] *CO2* *PO2* **6**

UNIT - III

5 a) Discuss the concept of cross tabulation in Pandas and explain how it is different from the Pivot table. Write a python program to demonstrate the same. *CO2* *PO2* **10**

b) Consider a dataset representing sales data for orders purchased. The dataset includes the following columns: *CO2* *PO2* **10**



```
     ord_no  purch_amt    ord_date  customer_id  salesman_id
0   70001.0     150.50  2012-10-05         3002       5002.0
1       NaN     270.65  2012-09-10         3001       5003.0
2   70002.0      65.26         NaN         3001       5001.0
3   70004.0     110.50  2012-08-17         3003          NaN
4       NaN     948.50  2012-09-10         3002       5002.0
5   70005.0    2400.60  2012-07-27         3001       5001.0
6       NaN    5760.00  2012-09-10         3001       5001.0
7   70010.0    1983.43  2012-10-10         3004          NaN
8   70003.0    2480.40  2012-10-10         3003       5003.0
9   70012.0     250.45  2012-06-27         3002       5002.0
10      NaN      75.29  2012-08-17         3001       5003.0
11  70013.0    3045.60  2012-04-25         3001          NaN
```

Write a Python program that performs the following tasks:

i. Display the original dataset with missing values.

ii. Analyze the missing values in the 'salesman_id' column and discuss possible reasons for their absence.

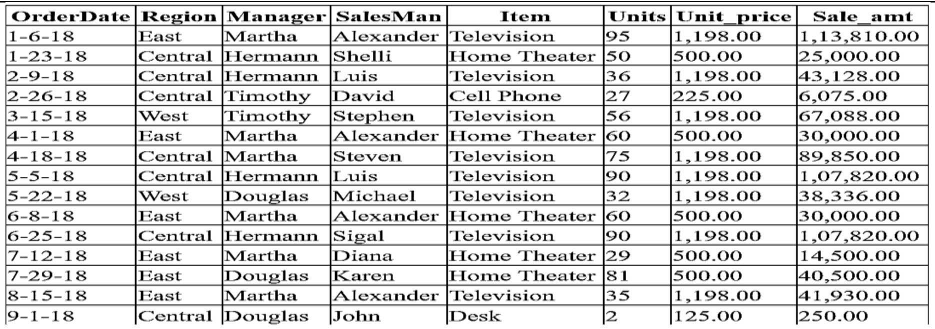
iii. Choose and implement the chosen method to fill in the missing values in the 'salesman_id' column.

iv. Display the dataset after filling in the missing values.

v. Calculate the total number of missing values in a DataFrame.

OR

6 a) Consider a dataset representing date wise sales data for various regions. The dataset includes the following columns *CO2* *PO2* **10**



OrderDate	Region	Manager	SalesMan	Item	Units	Unit_price	Sale_amt
1-6-18	East	Martha	Alexander	Television	95	1,198.00	1,13,810.00
1-23-18	Central	Hermann	Shelli	Home Theater	50	500.00	25,000.00
2-9-18	Central	Hermann	Luis	Television	36	1,198.00	43,128.00
2-26-18	Central	Timothy	David	Cell Phone	27	225.00	6,075.00
3-15-18	West	Timothy	Stephen	Television	56	1,198.00	67,088.00
4-1-18	East	Martha	Alexander	Home Theater	60	500.00	30,000.00
4-18-18	Central	Martha	Steven	Television	75	1,198.00	89,850.00
5-5-18	Central	Hermann	Luis	Television	90	1,198.00	1,07,820.00
5-22-18	West	Douglas	Michael	Television	32	1,198.00	38,336.00
6-8-18	East	Martha	Alexander	Home Theater	60	500.00	30,000.00
6-25-18	Central	Hermann	Sigal	Television	90	1,198.00	1,07,820.00
7-12-18	East	Martha	Diana	Home Theater	29	500.00	14,500.00
7-29-18	East	Douglas	Karen	Home Theater	81	500.00	40,500.00
8-15-18	East	Martha	Alexander	Television	35	1,198.00	41,930.00
9-1-18	Central	Douglas	John	Desk	2	125.00	250.00

Write a Python program using pandas that performs the following tasks:

i. Load the given dataset into a pandas DataFrame.

ii. Create a pivot table that shows the total sales for each product across
different regions.

iii. Calculate the average sales for each product.

iv. Identify the manager with highest sales_amt.

v. Determine the product that contributed the most to the sales in each region.

b) Explain the various measures of dispersion and classify the different skewness and Kurtosis measures available with examples? *CO2* *PO2* **10**

UNIT - IV

7 a) Differentiate between different types of Linear & Non-Linear scale explain each scale with an example. *CO3* *PO3* **10**

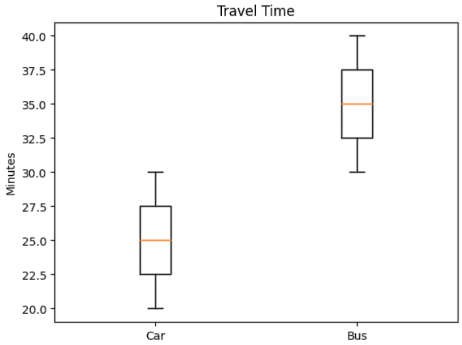
Plot the population densities (assuming your own value in crores) across Ten different states in India using Logarithmic scale display the output.

b) Consider the data below. Plot a suitable distribution by considering the data below. Pick up a suitable kernel smoothing function to plot values for the dataset below how this function is used for smoothing the data values. *CO3* *PO3* **10**

Age	Count
0-5	36
6-10	19
11-15	18
16-20	99
21-25	139
26-30	121
31-35	76

Age	Count
41-45	54
46-50	50
51-55	26
56-60	22
61-65	16
66-70	3
31-35	3

OR

<table>
<tr><td></td><td>8</td><td>a)</td><td>Consider the table below for different values of the variable Calculate the Cumulative distribution function. Plot the data points against cumulative probabilities obtained. Also explain the Quantile Quantile Distribution how the values are plotted using different intervals of Normal Distribution.
<table><tr><td>X</td><td>1</td><td>4</td><td>6</td><td>2</td><td>5</td><td>3</td></tr><tr><td>P(x)</td><td>0.1</td><td>0.3</td><td>0.02</td><td>0.2</td><td>0.02</td><td>0.2</td></tr></table></td><td><i>CO3</i></td><td><i>PO3</i></td><td><b>10</b></td></tr>
<tr><td></td><td></td><td>b)</td><td>Analyze the data below of Test Scores in students in different subject. Draw a suitable visualization technique which can represent this data.
<table><tr><th>Subject</th><th>Class A</th><th>Class B</th></tr><tr><td>Math</td><td>85</td><td>78</td></tr><tr><td>Science</td><td>90</td><td>84</td></tr><tr><td>English</td><td>88</td><td>86</td></tr></table></td><td><i>CO3</i></td><td><i>PO3</i></td><td><b>5</b></td></tr>
<tr><td></td><td></td><td>c)</td><td>Examine the following boxplot and answer the questions.


• Which mode of transportation has a higher median travel time?
• Calculate the interquartile range (IQR) for both Car and Bus.
• Identify if there are any outliers in the travel time of the Bus.
• Compare the variability in travel time between Car and Bus.</td><td><i>CO2</i></td><td><i>PO2</i></td><td><b>5</b></td></tr>
<tr><td></td><td></td><td></td><td><b>UNIT - V</b></td><td></td><td></td><td></td></tr>
<tr><td></td><td>9</td><td>a)</td><td>Create a suitable code to perform web scraping using URL http://www.geeksforgeeks.org and print the html documentation by using suitable library. Perform Web scraping also for the url http://example.com to print Xml documentation.</td><td><i>CO2</i></td><td><i>PO2</i></td><td><b>10</b></td></tr>
<tr><td></td><td></td><td>b)</td><td>Differentiate between serialization & deserialization in pandas. With an Example python code create a random of 100 numbers by storing the frame created in HDF5 binary format.</td><td><i>CO2</i></td><td><i>PO2</i></td><td><b>10</b></td></tr>
<tr><td></td><td></td><td></td><td><b>OR</b></td><td></td><td></td><td></td></tr>
<tr><td></td><td>10</td><td>a)</td><td>Explain the concept of hierarchical Indexing. Create a Multilevel Index for a random series of 12 numbers with corresponding row and column labels with a dimension of four cross three by writing a suitable python code. Display the output of python code.</td><td><i>CO3</i></td><td><i>PO3</i></td><td><b>8</b></td></tr>
<tr><td></td><td></td><td>b)</td><td>Perform stacking and unstacking operation by creating a suitable data frame of two cross three (2x3 Matrix). Index the dataframe by suitable column name as (a,b.c) and row names by person1 and person2.</td><td><i>CO3</i></td><td><i>PO3</i></td><td><b>6</b></td></tr>
<tr><td></td><td></td><td>c)</td><td>Perform the following operation by creating two two-dimensional arrays using numpy i) addition ii) cross-product iii) dot-product iv) subtraction. Display the result obtained from the above operation.</td><td><i>CO3</i></td><td><i>PO3</i></td><td><b>6</b></td></tr>
</table>