Sardar Patel Institute of Technology

Bhavan's Campus, Munshi Nagar, Andheri(West),Mumbai 400058-India

(An Autonomous Institute Affiliated to University of Mumbai)

End Semester Examination

Max. Marks: 100	**Duration:3 Hrs**
Class: SYMCA	**Semester: III**
Course Code:MC513	**Date: 20/12/23**
Course: Data Visualization and Analytics	**Time: 10 to 1**

Instructions:
  (1) All Questions are Compulsory.
  (2) Draw neat diagrams.
  (3) Assume suitable data if necessary.

No	Question	Max. Marks	CO	BL
Q1 A	What is Time Series Analysis? Explain different methods of Time series forecasting with example.	10	1	3
Q1 B	What are the different types of distributions? Explain different distributions with examples.	10	1	3
Q2 A	Which are the different libraries used in the python for Data Visualization? Explain in details.	10	3	3
Q2 B	What is Scatter plot? Write a python code to draw the Scatter Plot between two variables. **OR** What is box plot? Explain the five point summary contents of it. Also write a python code to draw a box plot of a particular variable.	10	3	3
Q3 A	The mean salaries of Data Scientists working in Chennai, India is calculated to be 7,00,000 INR with a standard deviation of 90,000 INR. The random variable salary of Data Scientists follows a normal distribution.What are the probabilities that a) What is the probability that a Data Scientist in Chennai has a salary more than 10,00,000 INR? b) What is the probability that a Data Scientist in Chennai has a salary between 6,00,000 & 9,00,000 INR? c) What is the probability that a Data Scientist in Chennai has a salary less than 4,00,000 INR? d) What is the probability that a Data Scientist in Chennai has a salary not more than 50,00,000 INR?	10	2	3
Q3 B	Explain the bivariate analysis for Categorical and continuous variables with examples and appropriate plots.	10	3	3
Q4 A	What is Power Query? Explain any two Text, Number and Date functions with example used in Power Query. **OR**	10	4	2

		What is Power BI? Describe its features, advantages, and disadvantages.			

Q4 B	Describe the points to be considered while designing an Information Dashboard.	10	5	2

Q5 A	**Dataset Details:** The honeyproduction.csv data set contails the following columns.	10	2	3

numcol: Number of honey producing colonies. Honey producing colonies are the maximum number of colonies from which honey was taken during the year. It is possible to take honey from colonies which did not survive the entire year

yieldpercol: Honey yield per colony. Unit is pounds

totalprod: Total production (numcol x yieldpercol). Unit is pounds

stocks: Refers to stocks held by producers. Unit is pounds

priceperlb: Refers to average price per pound based on expanded sales. Unit is dollars.

prodvalue: Value of production (totalprod x priceperlb). Unit is dollars.

Other useful information: Certain states are excluded every year (ex. CT) to avoid disclosing data for individual operations. Due to rounding, total colonies multiplied by total yield may not equal production. Also, summation of states will not equal U.S. level value of production.

Sample data is shown below:

Dataset Details: The honeyproduction.csv data set contails the following columns.

- numcol: Number of honey producing colonies. Honey producing colonies are the maximum number of colonies from which honey was taken during the year. It is possible to take honey from colonies which did not survive the entire year
- yieldpercol: Honey yield per colony. Unit is pounds
- totalprod: Total production (numcol x yieldpercol). Unit is pounds
- stocks: Refers to stocks held by producers. Unit is pounds
- priceperlb: Refers to average price per pound based on expanded sales. Unit is dollars.
- prodvalue: Value of production (totalprod x priceperlb). Unit is dollars.
- Other useful information: Certain states are excluded every year (ex. CT) to avoid disclosing data for individual operations. Due to rounding, total colonies multiplied by total yield may not equal production. Also, summation of states will not equal U.S. level value of production.

Sample data is shown below.

	state	numcol	yieldpercol	totalprod	stocks
0	AL	16000.0	71	136000.0	59000.0
1	AZ	55000.0	60	3300000.0	1485000.0
2	AR	53000.0	65	3445000.0	1688000.0
3	CA	450000.0	83	37350000.0	12326000.0
4	CO	27000.0	72	1944000.0	1594000.0

priceperlb	prodvalue	year
0.72	818000.0	1998
0.64	2112000.0	1998
0.59	2033000.0	1998
0.62	23157000.0	1998
0.70	1361000.0	1998

Write a python code to answer below questions based on the above dataset.
i) How many rows and columns does the data set have?¶
ii How many States are included in the dataset?
iii) Calculate the average production for each state across all years
iv) How many years data is provided in the dataset? And what is the starting and ending year?
v) Give the statistical summary of data.

Q5 B	i) Give some data visualization examples using following visual variables 1. shape 2. orientation 3. color 4. texture 5. value 6. size 7. Position 8. order ii) Arrange the steps in data visualization 1. Define a clear purpose. 2. Keep visualizations simple. 3. Make it actionable. 4. Provide context 5. Know your audience 6. Choose the right visual. 7. Make sure your visualizations are inclusive.	5+5	1	2