Reg. No. : | | | | | | | | | | | | |

Question Paper Code : 50898

B.E./B.Tech. DEGREE EXAMINATIONS, APRIL/MAY 2024.

Third/Fifth/Sixth Semester

Computer Science and Engineering

CS 3352 — FOUNDATIONS OF DATA SCIENCE

(Common to : Computer Science and Engineering (Artificial Intelligence and Machine Learning)/ Computer Science and Engineering (Cyber Security)/Computer and Communication Engineering/Electronics and Instrumentation Engineering/ Instrumentation and Control Engineering / Information Technology)

(Regulations 2021)

Time : Three hours Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. How missing values present in a dataset are treated during data analysis phase?

2. Identify and write down various data analytic challenges faced in the conventional system.

3. Will treating categorical variables as continuous variables result in a better predictive model? Justify your answer.

4. Issue: Feeding data which has variables correlated to one another is not a good statistical practice, since we are providing multiple weightage to the same type of data.

   Solution: Correlation Analysis.

   Show how such issues are prevented by correlation analysis technique. Justify with a small instance dataset.

5. State the purpose of adding additional quantitative and/or categorical explanatory variables to any developed linear regression model. Justify with an example.

6. Give an example of a data set with a non-Gaussian distribution.

7. Under what circumstances, the pivot_table( ) in pandas is used?

8. Using appropriate data visualization modules develop a python code snippet that generates a simple sinusoidal wave in an empty gridded axes?

9. Write a python code snippet that generates a time-series graph representing COVID-19 incidence cases for a particular week.

Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
7	18	9	44	2	5	89

10. Write a python code snippet that draws a histogram for the following list of positive numbers.

7	18	9	44	2	5	89	91	11	6	77	85	91	6	55

PART B — (5 × 13 = 65 marks)

11. (a) (i) Suppose there is a dataset having variables with missing values of more than 30%, how will you deal with such dataset? (6)

(ii) List down the various feature selection methods for selecting the right variables for building efficient predictive models. Explain about any two selection methods. (7)

Or

(b) (i) Explain Data Analytic life cycle. Brief about Time-Series Analysis. (6)

(ii) Outline the purpose of data cleansing. How missing and nullified data attributes are handled and modified during preprocessing stage? (7)

12. (a) (i) Indicate whether each of the following distributions is positively or negatively skewed. The distribution of

(1) Incomes of tax payers have a mean of $48,000 and a median of $43,000. (3)

(2) GPAs for all students at some college have a mean of 3.01 and a median of 3.20. (3)

(ii) During their first swim through a water maze, 15 laboratory rats made the following number of errors (blind alleyway entrances): 2, 17, 5, 3, 28, 7, 5, 8, 5, 6, 2, 12, 10, 4, 3.

(1) Find the mode, median, and mean for these data. (3)

(2) Without constructing a frequency distribution or graph, would it be possible to characterize the shape of this distribution as balanced, positively skewed, or negatively skewed? (4)

Or

(b) (i) Assume that SAT math scores approximate a normal curve with a mean of 500 and a standard deviation of 100.

Sketch a normal curve and shade in the target area(s) described by each of the following statements:

- More than 570 (2)
- Less than 515 (2)
- Between 520 and 540 (2)
- Convert to z scores and find the target areas specific to the above values. (1)

(ii) Assume that the burning times of electric light bulbs approximate a normal curve with a mean of 1200 hours and a standard deviation of 120 hours. If a large number of new lights are installed at the same time (possibly along a newly opened freeway), at what time will

- 1 percent fails? (2)
- 50 percent fail? (2)
- 95 percent fail? (2)

13. (a) (i) In Statistics, highlight the impact when the goodness of fit test score is low? (6)

(ii) Given the following dataset of employee, Using regression analysis, find the expected salary of an employee if the age is 45. (7)

Age	Salary
54	67000
42	43000
49	55000
57	71000
35	25000

Or

(b) (i) Define autocorrelation and how is it calculated? What does the negative correlation convey? (6)

(ii) What is the philosophy of Logistic regression? What kind of model it is? What does logistic Regression predict? Tabulate the cardinal differences of Linear and Logistic Regression. (7)

14. (a) Define Dictionary in Python. Do the following operations on dictionaries.

(i) Initialize two dictionaries (D_1 and D_2) with key and value pairs. (3)

(ii) Compare those two dictionaries with master key list 'M' and print the missing keys. (3)

(iii) Find keys that are in D_1 but NOT in D_2. (3)

(iv) Merge D_1 and D_2 and create D_3 using expressions. (4)

Or

(b) (i) How to create hierarchical data from the existing data frame? (6)

(ii) How to use group by with 2 columns in data set? Give a python code snippet. (7)

15. (a) Write a code snippet that projects our globe as a 2-D flat surface (using cylindrical project) and convey information about the location of any three major Indian cities in the map (using scatter plot). (13)

Or

(b) (i) Write a working code that performs a simple Gaussian process regression (GPR), using the Scikit-Learn API. (6)

(ii) Briefly explain about visualization with Seaborn. Give an example working code segment that represents a 2D kernel density plot for any data. (7)

PART C — (1 × 15 = 15 marks)

16. (a) Given a unsorted multi indexes that represents the distance between two cities, write a python code snippet using appropriate libraries to find the shortest distance between any two given cities. The following matrix representation can be used to create the data frame that can be served as an input for the prescribed program. (15)

	A	B	C	D	E
A	0	30	24	6	13
B	16	0	19	5	10
C	7	16	0	15	12
D	9	17	22	0	18
E	21	8	9	11	0

Or

(b) A URL Server wants to consolidate a history of websites visited by an user 'U'. Every visited website information is stored in a 2-tuple format *viz., (website_id, Duration_of_visit)* in the URL cache. Using split, apply and combine operations, devise a code snippet that consolidate the website history and find out the website whose duration of visit is maximum.

Example :

Input: [(4,2), (5,1), (4,3), (1,4), (7,3), (5,2), (1,1), (7,1)]

Output: [(4,5), (5,3), (1,5), (7,4)].

The website with key_id '1' has the max.duration of visit = 5. (15)

———————