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

Question Paper Code : 40918

B.E./B.Tech. DEGREE EXAMINATIONS, NOVEMBER/DECEMBER 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. Discern the difference between data analytics and data science.

2. Give an approach to solve any data analytics based project.

3. Suppose there is a dataset having variables with missing values of more than 30%, paraphrase about how to deal with such a dataset?

4. What does a "normal distribution" mean in statistics? Give an example. How to identify if the distribution is normal?

5. List down the various types of multiple regression.

6. List down the properties of correlation.

7. With appropriate syntax show how to create hierarchical data from the existing data frame.

8. Write a python code snippet that creates DataFrame objects for the following list of dictionaries.

   D_1 = [{'a': 1,'b': 2},{'b': 3,'c': 4}]

   D_2 = {'a': i,'b': (4 * i)/2.3} where i belongs to the range (0,4].

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
71	8	92	41	23	5	2

10. With suitable examples brief about the following indexing attributes: *Ioc*,*iloc*, and *ix*.

PART B — (5 × 13 = 65 marks)

11. (a) (i) Identify and list down various data analytic challenges faced in the conventional system. Explain about any two challenges in detail. (6)

(ii) Explain in brief about Exploratory Data Analysis. (7)

Or

(b) Explain in brief about the following measures adopted in describing the characteristics of data.

- measures of central tendency (5)
- measures of variability (4)
- frequency distribution (4)

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

(1) Incomes of taxpayers have a mean of Rs:12,000/- and a median of Rs: 13,450/-. (3)

(2) GPAs for all students at some college have a mean of 7.82 and a median of 8.21. (3)

(ii) During their first swim through a water maze, 15 laboratory rats made the following number of errors (blind alleyway entrances): 15, 3, 6, 12, 8, 1, 6, 4, 3, 2, 2, 1 and 10.

(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) What does a z-score reveal? How to interpret it? Assume that the grades on a Statistics mid semester exam at a college have a mean of $\mu = 63$, and a standard deviation of $\sigma = 3.0$. Advaith scored 65 on the exam. Find the z-score for Advaith's exam grade round to two decimal places. (6)

(ii) Assume that the burning times of electric light bulbs approximate a normal curve with a mean of 1152 hours and a standard deviation of 114 hours. If a large number of new lights are installed at the same time, at what time will

- 3 percent fails?
- 50 percent fail?
- 92 percent fail? (7)

13. (a) (i) In Statistics, what happens 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 48. (7)

Age	Salary
52	48500
46	46450
39	38750
41	41500
59	55500

Or

(b) (i) Explain in detail about fitting a Multiple Regression model. Is the Multiple Regression model a good fit? Outline the cardinal parameters to be considered while fitting the model. (6)

(ii) Define the terms correlation coefficient (r) and its square (r^2). What is the coefficient of determination (R^2)? When to use each of them and outline their differences. (7)

14. (a) (i) Write down a suitable example to drop Null values in a DataFrame using pandas.

(ii) How to fill null values in Pandas. Give a suitable example.

(iii) Give a code snippet to merge two Dataframes.

(iv) Explain in brief about slicing a Dataframe into multiple pieces in pandas. Give a suitable example. (3+3+3+4)

Or

(b) (i) Explain in brief about pandas MultiIndex. Give the example code snippet. (6)

(ii) Explain in brief about *one-to-one*, *many-to-one* and *many-to-many* joins in Pandas Data manipulation. Give sample working codes for the same. (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).

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) Assume that two datasets in the form of string arrays D_1 and D_2 are provided.

A string S_2 is a subset of string S_1 if every letter in S_2 occurs in S_1 including multiplicity. For example, "wrr" is a subset of "warrior" but is not a subset of "world". A string S_1 from D_1 is universal if for every string S_2 in D_2, S_2 is a subset of S_1. Return an array of all the universal strings in D_1. The expected output may be returned in any order.

Example :

Input: D_1 = ["amazon","apple","facebook","google","SoRoCoent"],
D_2 = ["e","o"]

Output: ["facebook","google","SoRoCoent"]

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 (GroupBy), 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 as 5.