Reg. No. : E N G G T R E E . C O M

Question Paper Code : 70072

B.E./B.Tech. DEGREE EXAMINATIONS, NOVEMBER/DECEMBER 2022.

Third Semester

Computer Science and Engineering

CS 3352 – FOUNDATIONS OF DATA SCIENCE

(Common to: Computer and Communication Engineering / Information Technology)

(Regulations 2021)

Time : Three hours Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. Define Data Science and Big Data.
2. List an overview of common errors in retrieving data and which cleansing solutions to be employed.
3. Classify the below list of data into their types: (a) ethnic group (b) age (c) family size (d) academic major (e) sexual preference (f) IQ score (g) net worth (dollars) (h) third-place finish (i) gender (j) temperature and write a brief note on them.
4. Differentiate discrete and continuous variables.
5. What is a percentile rank? Give an example.
6. Consider Helen sent 10 greeting cards to her friends and she received back 8 cards, what is the kind of relationship it is? Brief on it.
7. List the attributes of a Numpy array. Give an example for it.
8. Create a data frame with key and data pairs as Key-Data pair as A-10, B-20, A-40, C-5, B-10, C-10. Find the sum of each key and display the result as each key group.
9. What is the purpose of errorbar function in Matplotlib? Give an example.
10. Showcase 3-dimensional drawing in Matplotlib with corresponding Python Code.

PART B — (5 × 13 = 65 marks)

11. (a) Examine the different facets of data with the challenges in their processing.

Or

(b) Explore the various steps associated with data science process and explain any three steps of it with suitable diagrams and example.

12. (a) Demonstrate the different types of variables used in data analysis with an example for each.

Or

(b) The number of friends reported by Facebook users is summarized in the following frequency distribution.

FRIENDS	*f*
400 – above	2
350 – 399	5
300 – 349	12
250 – 299	17
200 – 249	23
150 – 199	49
100 – 149	27
50 – 99	29
0 – 49	36
Total	200

(i) What is the shape of this distribution?

(ii) Find the relative frequencies.

(iii) Find the approximate percentile rank of the interval 300–349.

(iv) Convert to a histogram.

(v) Why would it not be possible to convert to a stem and leaf display?

13. (a) (i) Categorize the different types of relationships using Scatter plots. (7)

(ii) Each of the following pairs represents the number of licensed drivers (X) and the number of cars (Y) for seven houses in my neighborhood:

Drivers (X)	Cars (Y)
5	4
5	3
2	2
2	2
3	2
1	1
2	2

(1) Construct a scatterplot to verify a lack of pronounced curvilinearity. (2)

(2) Determine the least squares equation for these data. (Remember, you will first have to calculate r, SSy and SSx) (2)

(3) Determine the standard error of estimate, Sy/x, given that $n = 7$. (2)

Or

(b) (i) In studies dating back over 100 years, it's well established that regression toward the mean occurs between the heights of fathers and the heights of their adult Sons.

Indicate whether the following statements are true or false.

(1) Sons of tall fathers will tend to be shorter than their fathers. (1)

(2) Sons of short fathers will tend to be taller than the mean for all sons. (1)

(3) Every son of a tall father will be shorter than his father. (1)

(4) Taken as a group, adult sons are shorter than their fathers. (1)

(5) Fathers of tall sons will tend to be taller than their sons. (1)

(6) Fathers of short sons will tend to be taller than their sons but shorter than the mean for all fathers. (1)

(ii) Interpret the value of r^2 in correlation based analysis. (7)

14. (a) Imagine you have a series of data that represents the amount of precipitation each day for a year in a given city. Load the daily rainfall statistics for the city of Chennai in 2021 which is given in a csv file Chennairainfall2021.csv using Pandas generate a histogram for rainy days, and find out the days that have high rainfall.

Or

(b) Consider that, an E-Commerce organization like Amazon, have different regions sales as NorthSales, SouthSales, WestSales, EastSales.csv files. They want to combine North and West region sales and South and East sales to find the aggregate sales of these collaborating regions Help them to do so using Python code.

15. (a) How text and image annotations are done using Python? Give an example of your own with appropriate Python code.

Or

(b) Appraise the following (i) Histograms (ii) Binnings (iii) Density with appropriate Python code.

PART C — (1 × 15 = 15 marks)

16. (a) Perform an exploratory data analysis for the following data with different types of plots:

The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.

Data attributes:-

Age of patient at the time of operation (numerical)

Patient's year of operation (year – 1900, numerical)

Number of positive axillary nodes detected (numerical)

Survival status (class attribute) 1 = the patient survived 5 years or longer, 2 = the patient died within 5 year

Or

(b) Assume that an r of – .80 describes the strong negative relationship between years of heavy smoking (X) and life expectancy (Y).

Assume, furthermore, that the distributions of heavy smoking and life expectancy each have the following means and sums of squares: 5 60 35 70 x y X Y SS SS

(i) Determine the least squares regression equation for predicting life expectancy from years of heavy smoking. (3)

(ii) Determine the standard error of estimate, Sy/x, assuming that the correlation of –.80 was based on $n = 50$ pairs of observations. (3)

(iii) Supply a rough interpretation of Sy/x. (3)

(iv) Predict the life expectancy for John, who has smoked heavily for 8 years. (3)

(v) Predict the life expectancy for Katie, who has never smoked heavily. (3)

Reg. No. : E N G G T R E E . C O M

Question Paper Code : 30118

B.E./B.Tech. DEGREE EXAMINATIONS, APRIL/MAY 2023.

Third Semester

Computer Science and Engineering

CS 3352 – FOUNDATIONS OF DATA SCIENCE

(Common to: Computer and Communication Engineering/Information Technology)

(Regulations 2021)

Time : Three hours Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. Outline the difference between structured data and unstructured data.

2. Define data mining.

3. Compare and contrast qualitative data and quantitative data with an example.

4. List the differences between a discrete variable and a continuous variable with an example.

5. What is the use of scatter plot?

6. Define correlation coefficient.

7. State the advantages of using Numpy arrays.

8. Outline the two types of Numpy's UFuncs.

9. State the two possible options in IPython notebook used to embed graphics directly in the notebook.

10. How plt.scatter function differs from plt.flot function?

PART B — (5 × 13 = 65 marks)

11. (a) Elaborate about the steps in the data science process with a diagram. (13)

Or

(b) What is a data warehouse? Outline the architecture of a data warehouse with a diagram. (13)

12. (a) (i) What is a frequency distribution? Customers who have purchased a particular product rated the usability of the product on a 10-point scale, ranging from 1 (poor) to 10 (excellent) as follows:

3	7	2	7	8
3	1	4	10	3
2	5	3	5	8
9	7	6	3	7
8	9	7	3	6

Construct a frequency distribution for the above data. (5)

(ii) What is relative frequency distribution? The GRE scores for a group of graduate school applicants are distributed as follows:

GRE Score	Frequency
725-749	1
700-724	3
675-699	14
650-774	30
625-649	34
600-624	42
575-599	30
550-574	27
525-549	13
500-524	4
475-499	2
Total	200

Explain the procedure to convert a frequency distribution into a relative frequency distribution and convert the data presented in the above table to a relative frequency distribution. Do not round numbers to two digits to the right of the decimal point. (8)

Or

(b) (i) What is Z-score? Outline the steps to obtain a Z-score. (7)

(ii) Express each of the following scores as a Z score: First, Mary's intelligence quotient is 135, given a mean of 100 and standard deviation 15. Second, Mary obtained a score of 470 in the Competitive Examination conducted in April 2022, given a mean of 500 and a standard deviation of 100. (6)

13. (a) Calculate the correlation coefficient for the heights 'in inches' of fathers' (x) and their son's (y) with the data presented below. (13)

x	66	68	68	70	71	72	72
y	68	70	69	72	72	72	74

Or

(b) The values of x and their corresponding values of y are presented below.

x	0.5	1.5	2.5	3.5	4.5	5.5	6.5
y	2.5	3.5	5.5	4.5	6.5	8.5	10.5

(i) Find the least square regression line $y = ax + b$ (9)

(ii) Estimate the value of y when $x = 10$. (4)

14. (a) What is an aggregate function? Elaborate about the aggregate functions in Numpy. (13)

Or

(b) (i) What is broadcasting? Explain the rules of broadcasting with an example. (7)

(ii) Elaborate about the mapping between Python operators and Pandas methods. (6)

15. (a) Explain about various visualization charts like line plots, scatter plots and histograms using Matplotlib with an example. (13)

Or

(b) Outline any two three-dimensional plotting in Matplotlib with an example. (13)

PART C — (1 × 15 = 15 marks)

16. (a) (i) What is mode? Can there be distributions with no mode or more than one mode? The owner of a new car conducts six gas mileage tests and obtains the following results, expressed in miles per gallon: 26.3, 28.7, 27.4, 26.6. 27.4, 26.9. Find the mode for these data. (5)

(ii) What is median? Outline the steps to find the median and find the median for the following scores: first, set of five scores 2, 8,2,7,6 and second, set of six scores 3,8,9,3, 1,8 with steps. (10)

Or

(b) Consider the following dataset with one response variable y and two predictor variables x_1 and x_2.

y	140	155	159	179	192	200	212	215
x_1	60	62	67	70	71	72	75	78
x_2	22	25	24	20	15	14	14	11

Fit a multiple linear regression model to this dataset. (15)

Reg. No. : E N G G T R E E . C O M

Question Paper Code : 20865

B.E./B.Tech. DEGREE EXAMINATIONS, NOVEMBER/DECEMBER 2023

Third/Fifth 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 and Information Technology)

(Regulations 2021)

Time : Three hours Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. What is Structured data?
2. Give an overview of common errors?
3. Explain the types of data.
4. Define median with example.
5. Define multiple regressions.
6. Define regression towards the mean.
7. What are the key properties of Pearson correlation coefficient?
8. Summarize some built – in Pandas aggregations?
9. Explain partial sort.
10. Give a summary about the comparison operators.

PART B — (5 × 13 = 65 marks)

11. (a) Explain the different facets of data with example.

Or

(b) Explain in detail about the cleansing, integrating, transforming data and build a model.

12. (a) (i) Explain Normal curve and Z-score. (6)

(ii) Using standard normal curve table, find the proportion of the total area identified with the following statements. (7)

(1) above a z score of 1.80

(2) between the mean and a z score of 1.65

(3) between z scores of 0 and –1.96

Or

(b) (i) Describe the types of variable. (6)

(ii) Suppose a hospital tested the age and body fat data for randomly selected adults with the following result :

Age	23	27	39	49	50	52	54	56	57	58	60
%fat	9.5	17.8	31.4	27.2	31.2	34.6	42.5	33.4	30.2	34.1	41

Draw the boxplots for age. (7)

13. (a) (i) Explain scatter plot. (6)

(ii) Describe range and variance. (7)

Or

(b) (i) Explain the correlation coefficient. (6)

(ii) Explain how the least squares equation which is used to minimize the total of all squared prediction errors with example. (7)

14. (a) Explain grouping in python with example.

Or

(b) Explain the following in python

(i) Data indexing (6)

(ii) Operation on missing data (7)

15. (a) Explain the different types of joins in python.

Or

(b) Explain various features of Matplotlib platform used for data visualization and illustrate its challenges.

PART C — (1 × 15 = 15 marks)

16. (a) Describe in detail about pivot table.

Or

(b) Find the following for the given data set :

Mean, Median, Mode, Variance, Standard Deviation and Skewness.

Marks	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
No. of Students	10	40	20	0	10	40	16	14

Reg. No. : | E | N | G | G | T | R | E | E | . | C | O | M |

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)