Reg. No. :

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

8

(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)