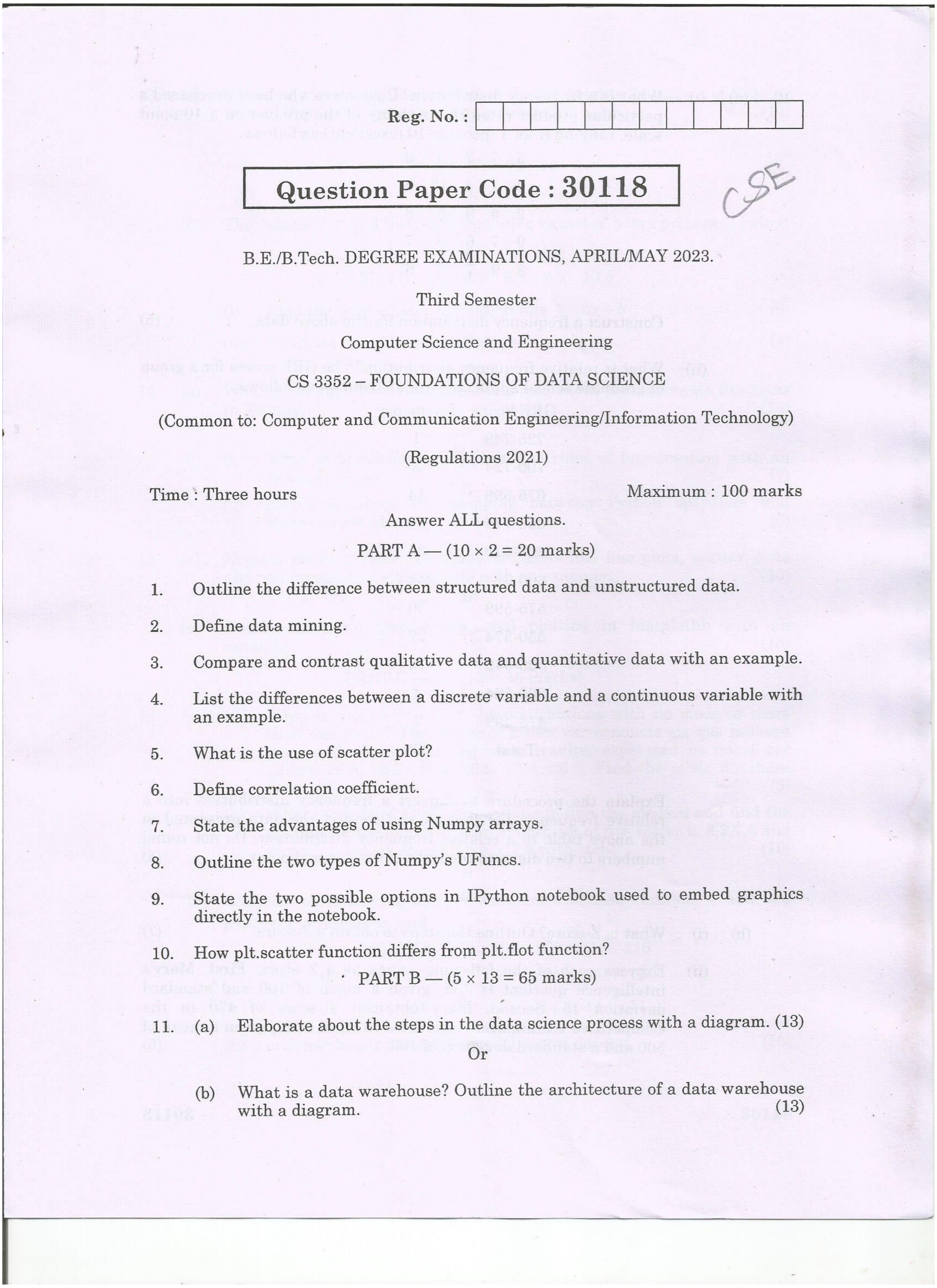


Reg. No. :

Question Paper Code : 30118

CSE

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.

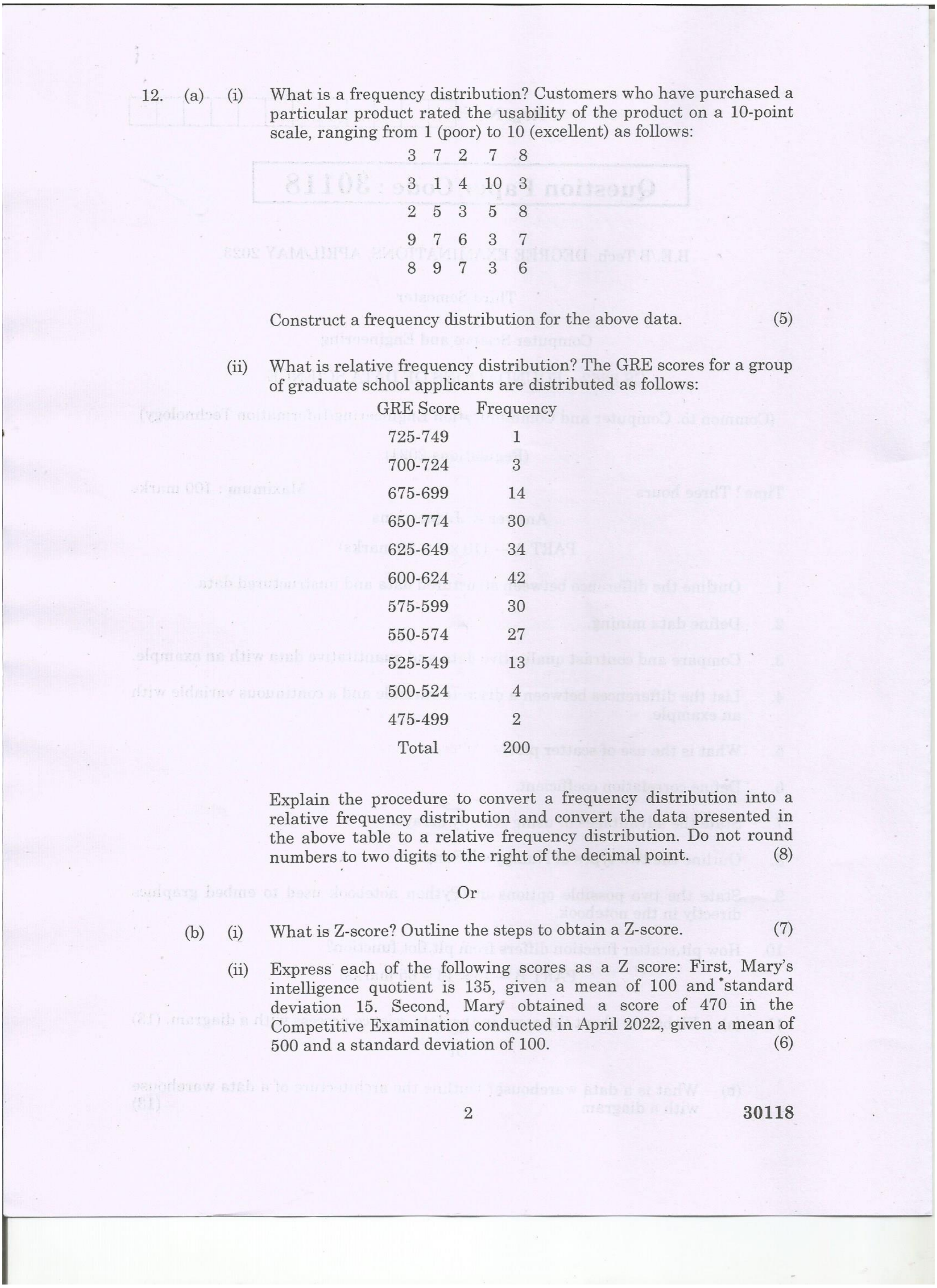
10. How plt.scatter function differs from plt.flot function?

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:

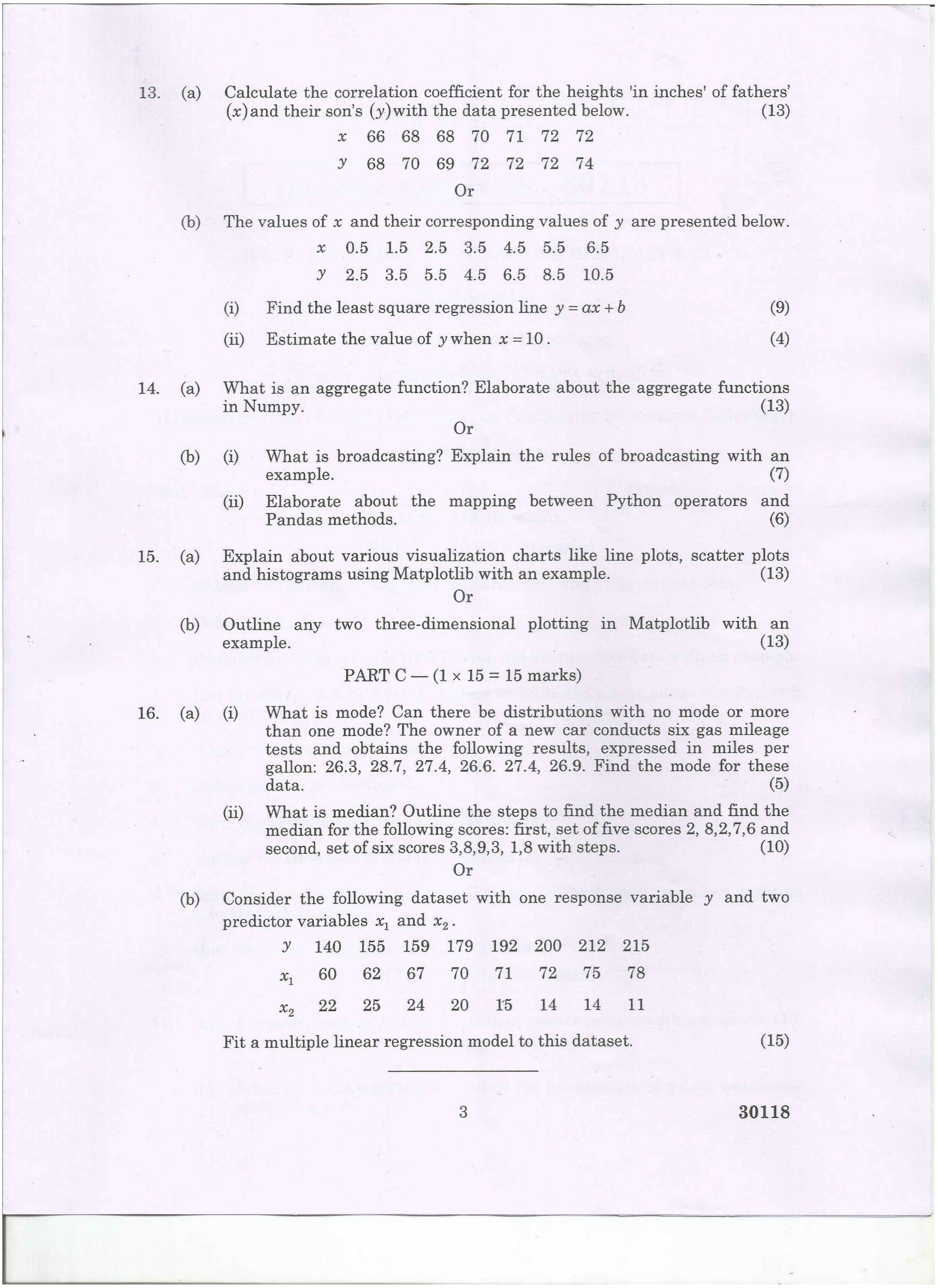
GRE Score	Frequency
725-749	1
700-724	3
675-699	14
650-774	30
625-649	34
600-624	42
575-599	30
550-574	27
525-549	13
500-524	4
475-499	2
Total	200

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)