

Reg. No. :

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Question Paper Code : 90920

B.E./B.Tech. DEGREE EXAMINATIONS, APRIL/MAY 2025.

Third 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/Information Technology)

(Regulations 2021)

Time : Three hours

Maximum : 100 marks

Answer ALL questions.

PART A — ($10 \times 2 = 20$ marks)

1. What is dummy variable?
2. Why is variable reduction significant?
3. Indicate whether the following quantitative observations are discrete or continuous.
 - (a) Litter of mice
 - (b) Cooking time for pasta
4. Find the mean for the following retirement ages: 60, 63, 45, 63, 65, 70, 55, 63, 60, 65, 63.
5. What is regression fallacy?
6. What is the significance of scatter plot?
7. What is broadcasting?

8. Write Python program to compute the component wise means of a list of vectors.
9. Write Python program to plot bar chart with an example.
10. What is histogram? Give an example.

PART B — (5 × 13 = 65 marks)

11. (a) (i) Explain various techniques to handle missing data. State their advantages and disadvantages. (8)
- (ii) Identify any outliers in each of the following sets of data collected from nine college students. (5)

SUMMER INCOME	AGE	FAMILY SIZE	GPA
\$6450	20	2	2.30
\$4,820	19	4	4.00
\$5,650	61	3	3.56
\$1,720	32	6	2.89
\$600	19	18	2.15
\$0	22	2	3.01
\$3,482	23	6	3.09
\$25,700	27	3	3.50
\$8,548	21	4	3.20

Or

- (b) Explain various stages in data science process with a diagram. (13)
12. (a) The IQ scores for a group of 35 high school dropouts are as follows:
 - (i) Construct a frequency distribution for grouped data.
 - (ii) Specify the real limits for the lowest class interval in this frequency distribution. (4+9)

91	85	84	79	80
87	96	75	86	104
95	71	105	90	77
123	80	100	93	108
98	69	99	95	90
110	109	94	100	103
112	90	90	98	89

Or

- (b) 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.

- (i) Find the mode, median, and mean for these data.
- (ii) Without constructing a frequency distribution or graph, would you characterize the shape of this distribution as balanced, positively skewed, or negatively skewed? (7+6)

13. (a) (i) After a group of college students attended a stress-reduction clinic, declines were observed in the anxiety scores of those who, prior to attending the clinic, had scored high on a test for anxiety.

- (1) Can this decline be attributed to the stress-reduction clinic? Explain your answer.
- (2) What type of study, if any, would permit valid conclusions about the effect of the stress reduction clinic? (6)

- (ii) Assume that an r of .30 describes the relationship between educational level and estimated hours spent reading each week. (7)

- (1) According to r^2 , what percent of the variability in weekly reading time is predictable from its relationship with educational level?
- (2) What percent of variability in weekly reading time is not predictable from this relationship?

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 squares:

$$\bar{X} = 5 \quad \bar{Y} = 60$$

$$SS_x = 35 \quad SS_y = 70$$

- (i) Determine the least squares regression equation for predicting life expectancy from years of heavy smoking. (3)
- (ii) Determine the standard error of estimate, $s_y | x$, assuming that the correlation of $-.80$ was based on $n = 50$ pairs of observations. (3)
- (iii) Predict the life expectancy for John, who has smoked heavily for 8 years. (3)
- (iv) Predict the life expectancy for Katie, who has never smoked heavily. (4)

14. (a) Write Python program to reads in lines of text and returns the ones that match a regular expression. (13)

Or

- (b) Write Python program that counts the words in its input and returns the most common ones. (13)

15. (a) (i) Write a note on density and contour plots. (6)
(ii) What is the requirement of Graphical Analysis? Explain. (7)

Or

- (b) What are the various types of graphs drawn using Matplotlib? Explain. (13)

PART C — ($1 \times 15 = 15$ marks)

16. (a) The body mass index (BMI) measures body size in people by dividing weight (in pounds) by the square of height (in inches) and then multiplying by a factor of 703. A BMI less than 18.5 is defined as underweight; between 18.5 to 24.9 is normal; between 25 and 29.9 is overweight; and 30 or more is obese. It is well established that Americans have become heavier during the last half century. Assume that the positively skewed distribution of BMIs for adult American males has a mean of 28 with a standard deviation of 4.

- (i) Would the median BMI score exceed, equal, or be exceeded by the mean BMI score of 28? (5)
(ii) What z score defines overweight? (5)
(iii) What z score defines obese? (5)

Or

- (b) Indicate whether the following generalizations suggest a positive or negative relationship. Also speculate about whether or not these generalizations reflect simple cause-effect relationships.

- (i) Preschool children who delay gratification (postponing eating one marshmallow to win two) subsequently receive higher teacher evaluations of adolescent competencies. (5)
(ii) College students who take longer to finish a test perform more poorly on that test. (5)
(iii) Infants who experience longer durations of breastfeeding score higher on IQ tests in later childhood. (5)