

Reg. No. :

4	2	2	1	2	2	1	0	4	0	2	6
---	---	---	---	---	---	---	---	---	---	---	---

Question Paper Code : 81159

B.E./B.Tech. DEGREE EXAMINATIONS, NOVEMBER/DECEMBER 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 Science and Engineering (Data Science)/Computer and Communication Engineering/Electrical and Computer Engineering/Information Technology)

(Regulations 2021)

Time : Three hours

Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. What is the main purpose of exploratory data analysis (EDA) in the data science process?
2. Differentiate between data mining and data warehousing.
3. What is the difference between qualitative and quantitative data?
4. Define standard (z) score and mention its significance.
5. What does the correlation coefficient (r) indicate about a relationship between two quantitative variables?
6. What is the purpose of the least squares regression line?
7. What is the purpose of using masks in NumPy arrays?
8. How does hierarchical indexing in Pandas help in data manipulation?
9. How do you create a simple line plot using Matplotlib?
10. How does Seaborn enhance data visualization compared to basic Matplotlib plots?

PART B — (5 × 13 = 65 marks)

11. (a) Explain the Data Science process in detail with each of its phases. Support your explanation with examples.

Or

- (b) Describe the benefits and uses of data science, elaborating on the various facets of data and how data mining and data warehousing support decision-making.

12. (a) Explain the different types of data and variables. Illustrate each type with appropriate examples.

Or

- (b) Discuss how data can be described using tables, graphs, measures of central tendency (averages) and measures of variability. Also, explain the concept of normal distribution and z-scores.

13. (a) Explain the concept of correlation using scatter plots and the correlation coefficient. Also, derive the computational formula for the correlation coefficient with a suitable example.

Or

- (b) Discuss regression analysis in detail, including the regression line, least squares method, standard error of estimate, interpretation of r^2 , and the concept of regression toward the mean. Explain with examples.

14. (a) Explain the basics of NumPy arrays and demonstrate array aggregations, computations, and advanced indexing techniques like boolean masks and fancy indexing with suitable examples.

Or

- (b) Discuss various data manipulation operations using Pandas, including data indexing and selection, handling missing data, aggregation, grouping, combining datasets, and creating pivot tables. Use real-world-like examples.

15. (a) Explain the different types of plots available in Matplotlib, including line plots, scatter plots, histograms, density and contour plots, and visualizing errors. Provide code examples for each.

Or

- (b) Discuss the role of Seaborn in data visualization, its integration with Matplotlib, and how it simplifies complex plots like categorical plots, pair plots, and heatmaps. Also, describe how geographic data can be visualized using Basemap and 3D plotting techniques in Matplotlib.

PART C — (1 × 15 = 15 marks)

16. (a) Explain the complete Data Science process from defining research goals to building applications. How are data types, variables, descriptive statistics, and regression models integrated into this process?

Or

- (b) Discuss how Python libraries such as NumPy, Pandas, Matplotlib, and Seaborn are used for data wrangling and visualization. Provide examples to illustrate key functionalities.

