Reg. No. : □□□□□□□□□□□□

Question Paper Code : 40029

B.E./B.Tech. DEGREE EXAMINATIONS, NOVEMBER/DECEMBER 2021.

Third Semester

Artificial Intelligence and Data Science

AD 8302 – FUNDAMENTALS OF DATA SCIENCE

(Regulations 2017)

Time : Three hours Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. Explain the concept of degrees of freedom in fundamental data analysis.
2. Differentiate between data and information with an illustrative example.
3. What is cluster analysis? Give any two techniques that can be used for cluster analysis.
4. State the various types of data distributions.
5. Define IQR. What is its significance?
6. Define Kurtosis and discuss its importance in understanding data characteristics.
7. State the different types of missing data with an example each.
8. Differentiate between univariate, bivariate and multivariate analysis.
9. What is dimensionality reduction? Why is it needed? Explain.
10. Explain Least-squares Regression.

PART B — (5 × 13 = 65 marks)

11. (a) What is Exploratory Data Analysis? Considering a dataset consisting of data on student performance in various courses of a semester, explain the major processes that are to be performed as part of the EDA process. (13)

Or

(b) Write a detailed note on the Descriptive data analysis techniques used in Data Science, for understanding the central tendency and shape of data. (13)

12. (a) Discuss the techniques that can be used to understand the variability characteristics of qualitative and ranked data. (13)

Or

(b) Discuss the need for applying smoothing to data during the Data cleaning step. Explain any two binning techniques used for smoothing discrete data, with relevant examples. Also, explain how binning can be analysed visually. (13)

13. (a) With relevant examples, explain how operations like aggregation, grouping, and data pivoting can be performed with Pandas. (13)

Or

(b) What is Hierarchical Indexing and why is it required? Explain how pandas can be used for creating hierarchical indexes. (13)

14. (a) Discuss the importance of Regression based analysis in Data Science. With a detailed example, show how Regression can be employed to provide important insights into the given data. (13)

Or

(b) Explain the various techniques that can be used for handling different levels of missing data. (13)

15. (a) Explain the process of using matplotlib for generating data visualization for different kinds of multidimensional datasets, to correctly capture their latent patterns for data analytics. (13)

Or

(b) Discuss the salient features of the Bokeh library, and how it can be used for interactive data visualization of large-scale datasets. (13)

PART C — (1 × 15 = 15 marks)

16. (a) A botanist claims that a particular strain of tomatoes developed by her lab is more disease resistant and is hence more durable. She conducts several trial studies with different batches of tomatoes and assigns a disease-resistance-quality score to each batch. A random sample of 30 batches of tomatoes is taken, and is found to have a mean quality score 168. Is there sufficient evidence to support the researcher's claim? The mean quality score of the entire tomato yield is standardized to 100, with a standard deviation of 22. Assume a normal distribution. Describe the process of proving your hypothesis in detail and derive relevant proof with valid reasoning. (15)

Or

(b) The age (*in years*) and systolic blood pressure of 10 apparently healthy adults is given. Analyze the correlation between age and blood pressure using Correlation Theory, and comment on your observations. What is the predicted blood pressure for a person of age 35 years? (15)

Age (x)	B.P (y)
20	120
43	128
63	141
29	126
53	134
31	128
58	136
46	132
58	140
70	144