Name: **Enrolment No:**	

UPES

End Semester Examination, December 2024

Course: Descriptive Analytics & Data Visualization **Semester: 5th**

Program: B.Tech. (Data Science) **Time : 03 hrs.**

Course Code: CSBA3018 **Max. Marks: 100**

No. of pages: 4

Instructions: Please attempt according to the provided time and given weightage.

SECTION A

(5Qx4M=20Marks)

Note: All Questions all compulsory from Section A

S.N.		Marks	CO
Q1	Categorize the various types of data and examine their specific ways of analysis in data analysis. Differentiate between univariate, bivariate, and multivariate analysis, and illustrate their applications with examples from real-world data.	**4**	**CO1**
Q2	Develop a conceptual model to represent the process of data generation, incorporating different data creation methods in the modern digital landscape. Assess the impact of data analytics in transforming raw data into actionable insights for business intelligence.	**4**	**CO1**
Q3	Investigate advanced visualization techniques such as chord diagrams, Sankey diagrams, and circular heat maps. Assess the effectiveness of these methods in overcoming the complexities of visualizing large and intricate datasets.	**4**	**CO2**
Q4	Evaluate the effectiveness of Python in data visualization tasks. Provide a detailed overview of key libraries such as Matplotlib and Seaborn, highlighting their unique capabilities and showcasing an example of how they can be used for data analysis.	**4**	**CO2**
Q5	Design a strategic approach to building a performance dashboard for a sales department. Select appropriate metrics and components using the Rayport-Jaworski framework and justify how these choices align with the department's business objectives.	**4**	**CO1**

SECTION B

(4Qx10M= 40 Marks)

Note: Q7, Q8, Q9 are compulsory. Attempt any one out of Q10, Q11.

Q7	An e-commerce company is looking to upgrade its Business Intelligence (BI) capabilities as part of a digital transformation initiative. Currently, the company relies heavily on spreadsheets and basic reporting tools for decision-making. Using the ASUG BI Maturity Model, assess the company's current BI maturity level. Provide a comprehensive plan to move from basic reporting to a fully integrated BI system, with a focus on strategic planning, system design, and development stages.	**10**	**CO1**

<table>
<tr>
<td>Q8</td>
<td>Critically analyze the concepts of covariance and correlation, explaining their significance in understanding the relationship between multiple variables. Using the provided dataset, calculate both the variance-covariance matrix and the correlation matrix. Evaluate the results and synthesize insights about the strength and direction of the relationships between the variables. Discuss the implications of these findings in the context of data analysis:
<table>
<tr><th>X</th><th>Y</th><th>Z</th></tr>
<tr><td>10</td><td>25</td><td>30</td></tr>
<tr><td>15</td><td>30</td><td>35</td></tr>
<tr><td>20</td><td>35</td><td>40</td></tr>
<tr><td>25</td><td>40</td><td>45</td></tr>
<tr><td>30</td><td>45</td><td>50</td></tr>
<tr><td>35</td><td>50</td><td>55</td></tr>
<tr><td>40</td><td>55</td><td>60</td></tr>
<tr><td>45</td><td>60</td><td>65</td></tr>
</table>
</td>
<td>10</td>
<td>CO3</td>
</tr>
<tr>
<td>Q9</td>
<td>A healthcare analyst is studying whether there is a significant difference in the average recovery times (in days) among three different treatment methods (Treatment X, Treatment Y, and Treatment Z). The recovery times from 30 patients (10 from each treatment method) are recorded as follows:

Treatment X: 12, 14, 13, 15, 16, 14, 13, 15, 12, 14
Treatment Y: 18, 19, 20, 17, 18, 19, 21, 20, 19, 18
Treatment Z: 11, 9, 10, 12, 11, 10, 9, 8, 10, 9

Perform a one-way ANOVA to test whether there is a significant difference in the average recovery times among the three treatment methods at the 5% significance level.
<ol type="a">
Formulate the null and alternative hypotheses.
Perform the calculations to determine the F-statistic.
Interpret the results and evaluate the finding based on F-statistic.

Given: tabulated $F_{0.05}$ for 2 and 27 d.f. in 5% significance level= 3.35
</td>
<td>10</td>
<td>CO3</td>
</tr>
<tr>
<td>Q10</td>
<td>Explain the Baye`s theorem in detail and Construct a probabilistic model using Bayes' Theorem to analyze the probabilistic model of following events:
There are two bags A and B. A contains n white and 2 black balls. B contains 2 white and n black balls. One of the two bags is selected at random, and two balls are drawn from it without replacement. If both the balls drawn are white and the probability that the bag A was used to draw the balls is 6/7, find the value of n.</td>
<td>10</td>
<td>CO2</td>
</tr>
<tr>
<td>Q11</td>
<td>Design a statistical framework to compare the effectiveness of two Fertilizer, X and Y, using a one-tailed t-test. Use the data below to perform the test and conclude at a 5% significance level.
<table>
<tr><td></td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td></tr>
<tr><td>X</td><td>50</td><td>52</td><td>54</td><td>48</td><td>55</td><td>53</td><td>50</td><td>49</td></tr>
<tr><td>Y</td><td>52</td><td>55</td><td>56</td><td>53</td><td>58</td><td>54</td><td>55</td><td>57</td></tr>
</table>
Perform an independent t-test to analyze whether Fertilizer X demonstrates a statistically significant improvement over Fertilizer Y. Critically interpret the results and recommend strategies for enhancing experimental validity and data-driven decision-making in similar studies?</td>
<td>10</td>
<td>CO4</td>
</tr>
</table>

	Given: Tabulated t0.05 for 14 degrees of freedom in 5% significance level equals to 2.15		
	SECTION-C **(2Qx20M=40 Marks)** **Note: Q11 is Compulsory. Attempt any one from Q12 and 13.**		
Q 12	In a linear regression scenario where sales (in units) are influenced by advertising spend in Social Media, analyze the following dataset: **TV** | **Radio** | **Social Media** | **Sales** 1000 | 300 | 150 | 45 1200 | 400 | 200 | 50 1100 | 350 | 180 | 48 1500 | 500 | 300 | 60 a) Develop a linear regression model to understand the relationship between advertising spend and sales. b) Derive the coefficients and formulate the regression equation. c) Predict the sales for the spending in Social Media $250 million. d) Evaluate the performance of the model by calculating R2 and interpreting its significance.	**20**	**CO4**
Q13	Critically analyze the role of statistical analysis in business intelligence and decision-making. Compare parametric and non-parametric methods with examples. Explain the type I and type II error with proper examples. Given the productivity scores before and after training: **Employee** | **Before** | **After** 1 | 45 | 50 2 | 40 | 44 3 | 42 | 48 4 | 38 | 42 Conduct a paired t-test to determine if the training program significantly improved productivity at a 5% significance level. Given: tabulated t0.05 for 26 d.f. in 5% significance level equals to 2.06.	**20**	**CO5**
Q14	Analyze the following defect data from a manufacturing company: **Category** | **Frequency** X | 40 Y | 30 Z | 25 W | 5 • Create a histogram to represent the defect distribution and calculate the mode. • Compute the probability of selecting a defective product from Category Y and discuss its significance for quality control. • Perform a Chi-square goodness-of-fit test to check if defect frequencies are uniformly distributed at a 5% significance level. Given: tabulate Chi-square value with 15 d.f in 5% significance level equals to 24.996	**20**	**CO5**