ourdiu.com | Shared by Md. Ferdous Foysal Hridoy

Daffodil International University

Faculty of Science & Information Technology
Department of Computer Science & Engineering
Midterm Examination, Fall 2026
Course Code: CSE328, Course Title: Introduction to Data Science
Level: 3 Term:3 Batch: 66

Time: 01:30 Hrs **Marks: 25**

Answer ALL Questions

[The figures in the right margin indicate the full marks and corresponding course outcomes. All portions of each question must be answered sequentially.]

1.	a)	A private hospital wants to predict which patients will be readmitted within 30 days of discharge, using admission records, laboratory results, pharmacy records and doctors' notes. The project team has listed its activities, but not in order: A) Deploy the model as an alert in the hospital's patient-management system B) Meet doctors to define "readmission" and decide what action a prediction should trigger C) Remove duplicate patient records and impute missing laboratory values D) Extract admission, laboratory and pharmacy records from three separate hospital databases E) Plot length-of-stay distributions and examine how laboratory values relate to readmission F) Train a classification model on historical admissions G) Test the model on unseen patients and find that it misses 40% of the patients who were actually readmitted Explain the correct sequence of these activities and map each one to its CRISP-DM phase. Then justify what the team must do after activity G instead of proceeding to A, and identify the phase(s) to which the project should loop back.	5	CO1
2.	a)	A health-insurance company is preparing the following patient dataset for analysis:	5	CO2

Patient ID	Birth Year	Body Temp (°C)	Pain Level (0–10)	No. of Visits	Income Bracket	Blood Pressure	Postal Code	Recovery Time (days)
P301	1975	38.4	7	3	20k–50k	145/95	1207	12.5
P302	1998	36.9	2	1	< 20k	115/75	1340	4.0

Apply your knowledge of variable types and measurement scales to classify each variable as qualitative or quantitative (discrete or continuous) and state its scale of measurement. For each variable, state the most appropriate measure of central tendency. Identify at least three variables that look numeric but would produce a meaningless result if an analyst calculated their mean.

3.		The Ministry of Labour wants to study occupational health among 400,000 garment workers employed in 2,000 factories across four industrial zones (Dhaka, Gazipur,	5	CO2

ourdiu.com | Shared by Md. Ferdous Foysal Hridoy

Narayanganj and Chattogram). A full census is not feasible, so four researchers propose different approaches:

i. Rafi proposes randomly selecting 20 factories and surveying every worker inside them.

ii. Nusrat proposes grouping workers by factory size small (30% of workers), medium (50%) and large (20%) and sampling from each group in proportion.

iii. Imran proposes choosing every 10th worker from each factory's payroll list. The list is ordered by production line; each line has exactly 10 workers, and the line supervisor is always listed first.

iv. Tania notes that workers with occupational illnesses often hide them for fear of losing their jobs. She proposes starting with workers known to a labour-rights NGO and asking them to refer colleagues with similar conditions.

a) Analyze the four proposals: name the sampling technique behind each one and identify the specific bias or weakness it carries in this context. Apply appropriate sampling techniques to design a practical multi-stage plan that selects a representative sample of 1,200 workers. Describe how participants would be selected at each stage, and explain where (if anywhere) Tania's approach can be justified despite its limitations.

4. An e-commerce company in Dhaka is choosing between two delivery partners, SwiftX and RapidY. Each partner handled 12 identical inter-district test deliveries. The delivery times (in hours) were: **CO3**

Delivery	1	2	3	4	5	6	7	8	9	10	11	12
SwiftX	22	24	23	25	21	24	26	23	22	25	24	71
RapidY	15	30	18	33	20	28	35	17	31	26	19	28

SwiftX's 12th delivery took place during a nationwide transport strike. Looking only at the averages, the logistics manager says: "RapidY is faster on average, so it is the better partner."

a) Calculate the mean, median and sample standard deviation of the delivery times for both partners. Analyze why the mean and the standard deviation lead to a different conclusion about speed and consistency than the median does, and identify the direction of skewness of each distribution. **3**

b) Construct side-by-side boxplots for the two partners. Using the Interquartile Range (IQR) method, mathematically analyze whether any delivery qualifies as a formal outlier. Compare the IQRs of the two partners and explain whether the manager's conclusion is statistically justified. Should the 71-hour delivery be removed from the analysis? Justify your answer. **3**

c) SwiftX advertises that its mean inter-district delivery time is 24 hours. The company suspects the true mean is longer and tests H_0: $\mu = 24$ versus H_a: $\mu > 24$ at a 5% significance level using all 12 SwiftX deliveries (critical value $t_{0.05,\,11} = 1.796$). **4**

Calculate the t-statistic and apply the decision rule. Analyze why a sample mean that is 3.5 hours above the claim still does not produce a significant result, which type of error (Type I or Type II) the company risks with this decision, and what this implies for the company's confidence in SwiftX's advertised delivery time.