BI – Unit 4 (Data Preparation) – END-SEM PYQ Answers

MAY-JUNE 2023

Q3a) Explain Data Exploration in detail with example. [7 marks]

Data exploration is the initial phase of data analysis where analysts examine a dataset to understand its structure, distributions, and key statistical properties before applying more sophisticated models. It is also called Exploratory Data Analysis (EDA). The goal is to discover patterns, spot anomalies, test hypotheses, and guide the selection of appropriate analytical techniques.

Data exploration is broadly divided into three types based on the number of variables examined simultaneously:

1. Univariate Analysis — Analyzing one variable at a time

Univariate analysis examines each feature in isolation to understand its distribution and statistical properties.

- For Categorical Attributes (e.g., Gender, City, Product Category):
  - Graphical: Bar charts and pie charts show the frequency or proportion of each category. Example: A bar chart of 'Product Category' shows Electronics accounts for 40% of records.
  - Tabular: Frequency tables list the count and percentage for each category value.
- For Numerical Attributes (e.g., Age, Salary, Sales):
  - Graphical: Histograms show value distribution; box plots reveal spread and outliers.
  - Measures of Central Tendency: Mean (average), Median (middle value), Mode (most frequent value). Example: Mean salary = Rs. 55,000; Median salary = Rs. 48,000 — the gap suggests a right-skewed distribution with some high earners.
  - Measures of Dispersion: Range, Variance, Standard Deviation, IQR (Interquartile Range). Example: High std. deviation in sales figures suggests wide variation across stores.
  - Outlier Identification: Using box plots or z-scores. A data point with z-score > 3 (i.e., more than 3 standard deviations from the mean) is typically flagged as an outlier.

2. Bivariate Analysis — Analyzing two variables together

Bivariate analysis examines the relationship between two variables to determine if they are correlated, dependent, or independent.

- For two Numerical variables: Scatter plots and Pearson Correlation Coefficient (r). Example: Scatter plot of 'Advertising Spend' vs. 'Revenue' shows a positive linear trend; r = 0.85 confirms strong positive correlation.
- For one Numerical + one Categorical variable: Box plots grouped by category, or grouped bar charts. Example: Salary distribution (box plot) for each Department shows the IT department has significantly higher median salary.
- For two Categorical variables: Contingency tables (cross-tabulation) and Chi-squared test. Example: A 2×2 contingency table of 'Gender' vs. 'Purchase Decision (Yes/No)' can reveal whether gender influences purchase behavior.

3. Multivariate Analysis — Analyzing three or more variables

Multivariate analysis examines complex relationships among multiple variables simultaneously.

- Graphical: Heatmaps of correlation matrices show pairwise correlations among all numerical features at a glance. Pair plots (scatter matrix) show every possible pair of variables in one grid.

- Correlation Matrix: A symmetric table where each cell (i,j) shows the correlation coefficient between feature i and feature j. Values close to +1 or -1 indicate strong relationships; values near 0 suggest independence.
- Dimensionality Reduction (PCA) is often the next step after multivariate exploration to reduce correlated features into fewer principal components.

Note: *Example dataset: A retail dataset with Age, Income, and Purchase Amount. Multivariate analysis reveals that high-income middle-aged customers have the highest purchase amounts — a three-variable insight impossible with univariate analysis alone.*

Q3b) Explain Data Transformation in detail with example. [5 marks]

Data transformation is the process of converting raw data from its original format or structure into a cleaner, more appropriate form suitable for analysis or machine learning. Raw data is rarely in the right shape — it may have different scales, skewed distributions, or redundant features.

Why is Data Transformation Needed?

- Different features have different scales (e.g., Age: 18–65; Income: 10,000–1,00,00,000). Many algorithms (KNN, SVM, gradient descent-based) are scale-sensitive and will be dominated by high-magnitude features without normalization.
- Skewed distributions violate assumptions of statistical models that expect normally distributed inputs.
- New informative features can sometimes be constructed from existing ones (feature engineering).

Key Transformation Techniques

- 1. Min-Max Normalization (Scaling to [0,1]): x' = (x - x_min) / (x_max - x_min). Example: Age values 20–60 become 0.0–1.0. Use when the feature must be bounded to a specific range. Sensitive to outliers.
- 2. Z-score Standardization: x' = (x - mean) / std_dev. Result has mean = 0, std. dev = 1. Example: Income of Rs. 55,000 with mean Rs. 50,000 and std. dev Rs. 10,000 → standardized value = 0.5. Preferred for algorithms assuming Gaussian distribution.
- 3. Log Transformation: x' = log(x). Used to reduce right-skew in distributions like income or sales figures. Example: Salary distribution is right-skewed; log(salary) produces a more symmetric, bell-shaped distribution.
- 4. Feature Extraction: Deriving new, more meaningful features from raw data. Example: From a datetime field 'Timestamp', extract 'Hour of Day', 'Day of Week', and 'Is Weekend' — all of which may be more predictive than the raw timestamp.
- 5. Encoding Categorical Variables: Convert text categories to numbers for ML algorithms. One-hot encoding creates binary columns for each category (e.g., 'City' → [Is_Pune, Is_Mumbai, Is_Nagpur]).

Q3c) Explain Data Validation: Incompleteness, Noise, and Inconsistency of Input Data Quality. [5 marks]

Before data can be used for analysis, it must be validated for quality issues. The three primary quality problems are incompleteness, noise, and inconsistency.

1. Incomplete Data

Incomplete data refers to records where one or more attribute values are missing (NULL values) or

entire records are absent.

- Causes: System failures during data collection, optional fields left blank by users, sensor malfunctions, data migration errors.
- Impact: Missing values can bias statistical results, cause errors in machine learning models, and lead to incorrect conclusions.
- Handling strategies: Ignore the record (if few missing values), fill with mean/median/mode (imputation), use predictive models to estimate missing values, or flag the value as 'Unknown' as a separate category.
- Example: A customer table where 30% of rows have a missing 'Annual Income' field — models trained on income will be biased.

2. Noise

Noisy data contains incorrect or distorted values — errors introduced during measurement, data entry, or transmission.

- Causes: Measurement instrument errors, manual data entry typos, transmission interference.
- Types: Random errors (a sensor randomly records wrong temperature readings) and systematic errors (a weighing scale that always reads 2 kg too high).
- Handling: Smoothing techniques like binning (replace individual values with bin averages), regression smoothing, or applying domain knowledge to set valid ranges and reject out-of-range values.
- Example: An age field containing value '250' — clearly erroneous. A purchase amount of '-500' in an e-commerce dataset is noise.

3. Inconsistency

Inconsistent data contains conflicting or contradictory information within the same dataset or across integrated datasets from multiple sources.

- Causes: Different naming conventions across systems (e.g., 'M' vs. 'Male' vs. '1' all meaning male gender), denormalized data in different tables being updated independently, integration of legacy systems that used different standards.
- Handling: Data standardization (define master reference values and map all variants to them), referential integrity enforcement, Master Data Management (MDM) systems.
- Example: 'Date of Birth' field says 2000-01-15 but 'Age' field says 35 — the two contradict each other. Another example: same customer appears with ID 'C001' in one table and 'CUST_001' in another.

NOV-DEC 2023

Q3a) Compute Mean, Median, and Mode for the given frequency distribution. [7 marks]

Given data:

Class	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50
Frequency	2	28	125	270	303	197	65	10

Step 1: Basic Setup

Total N = 2+28+125+270+303+197+65+10 = 1000. Midpoints (m) for each class: 12.5, 17.5, 22.5, 27.5,

32.5, 37.5, 42.5, 47.5.

Step 2: Mean

Mean = Σ(f × m) / N

Calculation: (2×12.5) + (28×17.5) + (125×22.5) + (270×27.5) + (303×32.5) + (197×37.5) + (65×42.5) + (10×47.5)

= 25 + 490 + 2812.5 + 7425 + 9847.5 + 7387.5 + 2762.5 + 475 = 31225

Mean = 31225 / 1000 = 31.225

Step 3: Median

Median class = class containing the N/2 = 500th value. Cumulative frequencies: 2, 30, 155, 425, 728... The 500th value falls in class 30-35 (cumulative becomes 728 after including this class).

Median = L + [(N/2 - CF) / f] × h, where L = 30 (lower boundary), CF = 425 (cumulative freq before median class), f = 303 (freq of median class), h = 5 (class width).

Median = 30 + [(500 - 425) / 303] × 5 = 30 + [75/303] × 5 = 30 + 1.2376 ≈ 31.24

Step 4: Mode

The modal class is the class with the highest frequency = 30-35 (f = 303).

Mode = L + [f1 - f0) / (2f1 - f0 - f2)] × h, where L = 30, f1 = 303, f0 = 270 (preceding class freq), f2 = 197 (succeeding class freq), h = 5.

Mode = 30 + [(303-270) / (2×303 - 270 - 197)] × 5 = 30 + [33 / (606-467)] × 5 = 30 + [33/139] × 5 = 30 + 1.187 ≈ 31.19

Result: Mean ≈ 31.23, Median ≈ 31.24, Mode ≈ 31.19. The three values are very close, indicating a nearly symmetrical distribution centred around 31.

Q3b) [REPEATED] What is Data Transformation? Explain in Detail. [5 marks]

Q3c) Explain Univariate, Bivariate, and Multivariate Analysis with examples and applications. [5 marks]

These three forms of analysis are already covered in detail under Q3a (May-June 2023) above. Here is a concise summary with additional application context.

Type	Variables	Key Methods	Real-World Application
Univariate	1	Mean, Median, Mode, Histograms, Box plots	Quality control: Analyze diameter distribution of manufactured parts to spot defects
Bivariate	2	Scatter plots, Correlation, Contingency table	Marketing: Does ad spend predict revenue? Correlation analysis reveals the relationship.
Multivariate	3+	Correlation matrix, PCA, Pair plots	Medical diagnosis: Analyze blood pressure, BMI, cholesterol and age together to predict heart disease risk.

Q4a) What is a Contingency Table? What is Marginal Distribution? Justify with example. [7 marks]

A contingency table (also called a cross-tabulation or cross-tab) is a matrix-format table used to display and summarize the frequency distribution of two or more categorical variables simultaneously. It is one of the primary tools for bivariate analysis of categorical data.

Structure

Consider two categorical variables: Gender (Male/Female) and Purchase Decision (Yes/No). The contingency table is:

Gender \ Purchase	Yes	No	Row Total
Male	60	40	100
Female	80	20	100
Column Total	140	60	200

Each cell shows the joint frequency — e.g., 60 males said Yes. The row and column totals are called marginals.

Marginal Distribution

Marginal distribution refers to the frequency or probability distribution of one variable ignoring the other variable — found by summing across rows or columns.

- Row marginal (Gender): Males = 100 (50%), Females = 100 (50%). This is the distribution of gender alone.
- Column marginal (Purchase Decision): Yes = 140 (70%), No = 60 (30%). This is the distribution of purchase decision alone.

Marginal distributions tell you about each variable in isolation; the contingency table tells you about their joint behavior and potential dependency.

Interpretation and Chi-Squared Test

From the example, 80% of females purchased (80/100) vs. 60% of males (60/100). This difference suggests gender may influence purchase decisions. A Chi-squared test quantifies whether this difference is statistically significant or due to random chance.

Q4b) [REPEATED] Explain Data Validation: Incompleteness, Noise, Inconsistency. [5 marks]

Q4c) Explain Data Reduction Techniques: Sampling, Feature Selection, Principal Component Analysis. [5 marks]

Data reduction techniques reduce the volume of data while preserving the information most relevant for analysis. Working with smaller, more focused datasets speeds up algorithms and reduces overfitting.

1. Sampling

Sampling selects a representative subset of records from a large dataset rather than processing all of them.

- Random sampling: Every record has an equal probability of being selected. Simple and unbiased but may miss rare classes.

- Stratified sampling: The dataset is divided into strata (e.g., by income group), and samples are drawn proportionally from each stratum. Ensures rare but important groups are represented.
- Example: A bank has 1 million customer records. Sampling 10,000 records (1%) for model development dramatically reduces training time while preserving the statistical properties of the full dataset.

2. Feature Selection

Feature selection identifies and retains only the most relevant input variables (features) for a model, discarding redundant or irrelevant ones.

- Filter methods: Rank features by statistical scores (correlation with target, mutual information, chi-squared statistic) and keep the top-k. Fast but does not account for feature interactions.
- Wrapper methods: Train the model with different feature subsets and select the subset that gives the best validation performance (e.g., Recursive Feature Elimination — RFE). Slow but more accurate.
- Embedded methods: Feature selection happens during model training — e.g., LASSO regression shrinks irrelevant feature coefficients to zero.
- Example: A customer churn prediction model with 50 features uses RFE to identify that only 12 features (recency, frequency, monetary value, service calls, etc.) are necessary.

3. Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique that transforms correlated features into a smaller set of uncorrelated variables called principal components. Each principal component is a linear combination of the original features, ordered by the amount of variance it explains.

- Step 1: Standardize the data so all features have mean = 0 and variance = 1.
- Step 2: Compute the covariance matrix of the standardized features.
- Step 3: Compute eigenvectors and eigenvalues of the covariance matrix.
- Step 4: Sort eigenvectors by their eigenvalues (descending). The first eigenvector (PC1) explains the most variance.
- Step 5: Project the original data onto the top k eigenvectors to get the reduced dataset.
- Example: A dataset with 100 correlated financial features (revenue, profit, operating costs, etc.) may be reduced to 10 principal components that explain 95% of the total variance, making it suitable for clustering or visualization.

Note: *PCA does not select original features — it creates new ones. Feature Selection retains original features. Both reduce dimensionality but serve different purposes and have different interpretability implications.*

MAY-JUNE 2024

Q3a) Discuss the need for Data Pre-processing and any 2 techniques used. [6 marks]

Need for Data Pre-processing

Real-world data collected from business systems, sensors, surveys, and logs is almost never ready for direct analysis. Studies suggest that data scientists spend 60–80% of their time on data preparation. The reasons pre-processing is essential include:

- Incompleteness: Missing values in critical attributes can skew model outcomes.
- Noisiness: Erroneous values (typos, sensor errors) introduce false patterns.

- Inconsistency: Data integrated from multiple sources often has conflicting representations.
- Irrelevance: Datasets often include features with no predictive value, which waste computational resources and confuse models.
- Wrong scale: Features with vastly different numerical ranges cause scale-sensitive algorithms to be dominated by high-magnitude features.
- Without pre-processing, any model built on raw data will produce unreliable, biased, or incorrect results — 'Garbage in, garbage out.'

Technique 1: Data Cleaning (Handling Missing Values)

Data cleaning corrects or removes records with quality problems. For missing values specifically:

- Deletion: Remove rows where critical values are missing. Suitable only when the proportion of missing data is very small (<5%) to avoid losing information.
- Mean/Median Imputation: Replace missing numerical values with the column mean (for symmetrically distributed data) or median (for skewed data). Example: Replace missing 'Age' values with median age = 35.
- Mode Imputation: Replace missing categorical values with the most frequent category. Example: Replace missing 'City' with 'Pune' if it is the most common city.
- Predictive Imputation: Use a regression or KNN model to predict the missing value from other features. More accurate but computationally expensive.

Technique 2: Data Normalization / Standardization

Already detailed under Q3b (May-June 2023). The key techniques are min-max normalization and z-score standardization. Normalization is critical before applying distance-based algorithms (KNN, K-means) or gradient descent optimization.

Q3b) [REPEATED] What is Data Transformation? Why is it needed? Explain at least 3 techniques. [6 marks]

Q3c) What is Data Reduction? Explain Dimensionality Reduction and Data Compression [6 marks]

What is Data Reduction?

Data reduction refers to the process of obtaining a reduced representation of a dataset that is much smaller in volume but closely maintains the integrity and information content of the original data. The reduced data requires less storage and less computation time for mining.

The goal is to allow data mining and machine learning algorithms to run efficiently without significant loss of analytical power.

1. Dimensionality Reduction

Dimensionality reduction reduces the number of features (columns / attributes) in a dataset. High-dimensional data suffers from the 'curse of dimensionality' — as dimensions increase, the data becomes increasingly sparse and distance measures become unreliable.

- Feature Selection: Choose the most informative subset of original features. Retains interpretability.
- Principal Component Analysis (PCA): Transform correlated features into a smaller number of uncorrelated principal components. (Detailed under Q4c Nov-Dec 2023.)
- Linear Discriminant Analysis (LDA): Similar to PCA but supervised — projects data to maximize

class separability.
- Benefits: Reduces overfitting, speeds up training, enables 2D/3D visualization of high-dimensional data.

2. Data Compression

Data compression reduces the storage size of a dataset, either losslessly (exact reconstruction possible) or lossily (some information is sacrificed for greater compression).

- Lossless Compression: All original data can be reconstructed perfectly. Example: Run-Length Encoding (RLE) for categorical data — '5×Male' instead of 'Male, Male, Male, Male, Male'. Used in database compression and backup.
- Lossy Compression: Some data detail is lost, but the overall analytical value is preserved. Example: Wavelet transforms reduce signal data while keeping dominant frequency patterns. JPEG image compression in image analytics.
- Numerosity Reduction: Replace the full dataset with a compact representation — histograms (approximate distribution), sampling (a representative subset), or parametric models (store model parameters rather than raw data).
- Example: A time-series dataset of stock prices at millisecond intervals (millions of rows) can be compressed to hourly or daily OHLC (Open-High-Low-Close) summaries without losing the patterns needed for trend analysis.

Note: *Dimensionality reduction operates on the column axis (fewer features); numerosity reduction operates on the row axis (fewer records); compression reduces storage footprint.*

Q4a) Define Dirty Data. What are the Reasons for Dirty Data? [6 marks]

Definition

Dirty data (also called low-quality data or bad data) refers to data that is inaccurate, incomplete, inconsistent, improperly formatted, or otherwise unsuitable for reliable analysis. Dirty data produces misleading results — a model trained on dirty data will make poor predictions even with sophisticated algorithms.

Reasons for Dirty Data

- Data entry errors: Human operators make typos or use non-standard formats (e.g., typing 'Pune' as 'Pnue', or entering a phone number in a name field).
- Integration from multiple sources: When data is merged from different systems (ERP, CRM, legacy databases), different conventions cause conflicts — e.g., date format MM/DD/YYYY in one system and DD-MM-YYYY in another.
- Incomplete records: Mandatory fields left blank during collection, optional survey questions skipped by respondents.
- System or hardware failures: Network interruptions during data transfer cause truncated or corrupt records. Sensor failures produce erroneous readings.
- Deliberate falsification: Users provide false information (fake birthdates, fake email addresses) to bypass validation.
- Outdated information: Stale data that was once accurate — a customer changes address but the database isn't updated, resulting in deliveries to wrong locations.
- Lack of data governance: No standards enforced for data formats, naming conventions, or mandatory fields, resulting in free-form, unstructured entries.
- Duplication: The same real-world entity appears multiple times with slightly different attributes

(e.g., same customer registered as 'Raj Patel' and 'Rajesh Patel').

Q4b) Explain the Working of Binning with a Suitable Example. [6 marks]

Binning (also called discretization) is a data pre-processing technique that converts continuous numerical data into discrete categorical bins or intervals. It is used to reduce the effect of minor observation errors, smooth out noisy data, and prepare continuous features for algorithms that prefer discrete inputs.

Why Binning?

- Smoothing: Outliers and minor errors within a bin are replaced by a single representative value, reducing noise.
- Discretization: Decision trees and association rule mining often work better with discrete categories than with continuous values.
- Generalization: Converting specific ages (23, 24, 25) into a group ('Young Adults: 20-30') reduces overfitting.

Types of Binning

- Equal-width binning: Each bin covers the same numerical range. If the data ranges from 0 to 100 and you create 5 bins, each bin covers 20 units: [0-20), [20-40), [40-60), [60-80), [80-100].
- Equal-frequency binning: Each bin contains approximately the same number of data points. Handles skewed distributions better.

Worked Example

Given a set of age values: 4, 8, 15, 21, 21, 24, 25, 28, 34, 36, 38, 42. Divide into 3 equal-width bins: Bin 1 [4-18): {4, 8, 15}, Bin 2 [18-32): {21, 21, 24, 25, 28}, Bin 3 [32-46): {34, 36, 38, 42}.

Smoothing by bin means: Bin 1 mean = (4+8+15)/3 = 9, Bin 2 mean = (21+21+24+25+28)/5 = 23.8, Bin 3 mean = (34+36+38+42)/4 = 37.5.

After smoothing: Bin 1 → {9, 9, 9}, Bin 2 → {24, 24, 24, 24, 24}, Bin 3 → {38, 38, 38, 38}. The individual noise in each bin is eliminated and replaced by the bin's representative value.

Smoothing by bin boundaries: Each value in a bin is replaced by the nearest boundary value. For Bin 1 [4,15]: {4→4, 8→4, 15→15}. (Points closer to the lower boundary become the lower boundary value and vice versa.)

Q4c) What is Bivariate Analysis? Why is it important? Discuss the different types with examples. [6 marks]

Additional Concepts & Quick Reference

Data Quality Dimensions — Complete Summary

Quality Issue	Definition	Example	Solution
Incompleteness	Missing values or records	Customer age = NULL	Imputation or deletion
Noise	Incorrect, distorted values	Age = 250, Amount = -500	Binning, regression smoothing
Inconsistency	Conflicting values	DOB says 2000, Age	Standardization, MDM

Quality Issue	Definition	Example	Solution
	within/across data	says 35	
Duplication	Same entity recorded multiple times	Same customer, two IDs	Deduplication, record linkage
Inaccuracy	Correct format but wrong value	Wrong phone number entered	Cross-validation with trusted source

PCA — Intuitive Explanation

Imagine you have data points scattered in 3D space (three features). PCA finds the direction in which the data varies the most (PC1), then the direction of second-most variance perpendicular to PC1 (PC2), and so on. If 90% of the total variance is captured by PC1 and PC2, you can project all your 3D data points onto the 2D plane defined by PC1 and PC2 with minimal information loss. This makes 3D data visualizable in 2D and reduces computational cost significantly.

Measures of Central Tendency and Dispersion — Quick Reference

Measure	Formula / Definition	When to Use
Mean	Sum of all values / Count	Symmetrically distributed, no extreme outliers
Median	Middle value when sorted	Skewed distributions or when outliers are present
Mode	Most frequently occurring value	Categorical data or multi-modal numerical data
Variance	Avg. of squared deviations from mean	Measuring spread; basis for std. deviation
Std. Deviation	Square root of variance	Expressing spread in original units
IQR	Q3 - Q1 (middle 50% range)	Robust measure of spread; used in box plots for outlier detection

NOV-DEC 2025

Q3a) Explain Data Reduction in detail with example. [7]

Data reduction is the process of obtaining a compact representation of a dataset that produces the same (or nearly the same) analytical results as the original data. As datasets grow to terabytes and beyond, full-scale processing becomes computationally prohibitive — data reduction makes mining feasible without sacrificing meaningful accuracy.

There are four major categories of data reduction, each operating on a different aspect of the data:

1. Dimensionality Reduction — Reducing the Number of Features

High-dimensional data suffers from the 'curse of dimensionality': as the number of features grows, the data becomes increasingly sparse, distance measures become unreliable, and models overfit.

Dimensionality reduction addresses this by keeping only the most informative features.

- Feature Selection: Identifies and retains a subset of the original features that carry the most predictive value, discarding redundant or irrelevant ones. Methods include filter approaches (rank by correlation or mutual information), wrapper approaches (train models on subsets and evaluate), and embedded approaches (LASSO regression, which shrinks irrelevant coefficients to zero during training).
- Principal Component Analysis (PCA): Rather than selecting features, PCA transforms correlated features into a smaller number of uncorrelated principal components. Each component is a linear combination of original features, and they are ordered by how much variance they explain. Example: a dataset with 50 correlated financial metrics may be reduced to 8 principal components explaining 95% of total variance.
- Linear Discriminant Analysis (LDA): A supervised variant that projects features to maximize class separability rather than variance — used when classification is the end goal.

2. Numerosity Reduction — Reducing the Number of Records

Rather than reducing columns, numerosity reduction reduces rows — representing the data with a smaller set of records or a parametric model.

- Sampling: Select a statistically representative subset of records. Random sampling gives every record an equal chance; stratified sampling ensures proportional representation of each class (critical when classes are imbalanced). Example: a bank with 5 million customer records samples 50,000 (1%) for model development. If properly stratified (e.g., 5% defaulters preserved), the sample retains the population's statistical character.
- Histograms: Partition numerical data into bins and record only the bin boundaries and counts — replacing millions of individual values with a compact frequency distribution. This is a lossy but space-efficient approximation.
- Clustering and Centroid Representation: Cluster the data and represent each cluster by its centroid. The entire cluster's data is replaced by a single representative point. Useful when the data has natural groupings.
- Parametric Models: If a dataset fits a known distribution (e.g., Gaussian), store only the distribution parameters (mean and standard deviation) rather than the raw data — a dramatic reduction from millions of records to two numbers.

3. Data Compression — Reducing Storage Size

Data compression reduces the physical storage footprint of a dataset.

- Lossless Compression: The original data can be perfectly reconstructed. Run-Length Encoding (RLE) is a simple example — a sequence 'Male, Male, Male, Female, Female' becomes '3×Male, 2×Female'. Used in database storage, backup systems, and columnar formats like Parquet.
- Lossy Compression: Some information is sacrificed for greater compression, but the overall analytical value is preserved. Wavelet transforms reduce signal data while retaining dominant frequency components. JPEG compression of images in visual analytics pipelines.

4. Data Discretization — Reducing Value Range Granularity

Discretization converts continuous numerical attributes into discrete categorical ones, reducing the number of distinct values while preserving meaningful distinctions. This is covered in detail under Q4b below.

Full Example: Reducing an E-Commerce Dataset

An online retailer has a dataset with 10 million customer transactions, 120 attributes per record, and raw timestamp data at millisecond granularity. Data reduction might proceed as: (1) Feature Selection — a correlation analysis reduces 120 attributes to the 20 most relevant for churn prediction. (2) Sampling — 100,000 records are sampled with stratification to preserve the 3% churn rate. (3) Discretization — raw purchase amount (continuous) is binned into Low/Medium/High categories. (4) Compression — the resulting clean dataset is stored in columnar Parquet format with LZ4 lossless compression, reducing storage by 70%.

> **Note:** *Data reduction is always applied after data cleaning, not before. Reducing dirty data amplifies the errors rather than removing them.*

Q3b) [REPEATED] What is Data Transformation? Explain the Data Transformation Process in detail. [5]

Q3c) [REPEATED] Explain Data Validation: Incompleteness, Noise, and Inconsistency. [5]

Q4a) [REPEATED] Contingency Table and Marginal Distribution. [7]

Q4b) Short Notes on Data Discretization. [5]

Data discretization is a data pre-processing technique that converts continuous numerical attributes into discrete, categorical intervals (bins or ranges). It reduces the number of distinct values an attribute can take, making data easier to work with for rule-based algorithms, association rule mining, and decision tree learning that prefer categorical inputs over fine-grained continuous values.

Why Discretization is Needed

- Association rule mining algorithms (Apriori) operate on categorical items — a continuous attribute like 'age = 23.7 years' cannot appear in an association rule, but 'age = Young Adult (20-30)' can.
- Decision trees can handle continuous attributes, but discretizing them in advance speeds up the tree-building process.
- Discretization achieves noise smoothing — minor measurement errors within a bin are absorbed into the bin's representative value rather than introducing spurious variation.
- It can improve model generalizability by preventing models from memorizing irrelevant fine-grained distinctions that do not hold in new data.

Types of Discretization

- Equal-Width (Equal-Interval) Binning: Divides the value range [min, max] into k bins of equal width. Width = (max - min) / k. Example: Income ranging from Rs. 10,000 to Rs. 1,00,000, divided into 3 equal-width bins: Low [10,000-40,000), Medium [40,000-70,000), High [70,000-1,00,000]. Simple to implement but sensitive to outliers — a single extreme value stretches the range.
- Equal-Frequency (Equal-Depth / Quantile) Binning: Ensures each bin contains approximately the same number of records. The bin boundaries are set at quantile points. This handles skewed distributions better than equal-width because all bins are equally populated. Example: 100 records divided into 4 equal-frequency bins of 25 records each — the boundaries adjust to the data distribution.
- Entropy-Based Discretization: A supervised method that uses the class label to find cut points

that maximize information gain. More accurate for classification tasks but requires labels. Used in algorithms like C4.5.

- Custom / Domain-Driven Discretization: Business experts define bins based on domain meaning — e.g., credit scores: Poor (<580), Fair (580-669), Good (670-739), Very Good (740-799), Exceptional (800+). This produces the most interpretable results.

Smoothing Methods Applied After Binning

- Bin mean smoothing: All values in a bin are replaced by the mean of the bin. Reduces noise by averaging out individual variations within each bin.
- Bin median smoothing: All values are replaced by the bin median. More robust to outliers within the bin.
- Bin boundary smoothing: Each value is replaced by the nearest bin boundary (lower or upper boundary). Preserves edge values while merging interior values to a boundary.

Worked Example

Given ages: 22, 25, 27, 30, 35, 38, 42, 45, 50. Discretized into 3 equal-width bins across the range [22, 50], width = (50-22)/3 ≈ 9.3: Bin 1 [22-31): {22, 25, 27, 30} → label 'Young'. Bin 2 [31-40): {35, 38} → label 'Middle'. Bin 3 [40-50]: {42, 45, 50} → label 'Senior'. After bin-mean smoothing: Bin 1 values all become 26, Bin 2 values become 36.5, Bin 3 values become 45.7.

> **Note:** *Discretization is sometimes called 'binning' when referring specifically to the equal-width or equal-frequency approaches. The term 'discretization' is broader and encompasses all methods that convert continuous to categorical, including entropy-based and domain-driven approaches.*

Q4c) [REPEATED] Data Reduction Techniques: Sampling, Feature Selection, PCA. [5]

MAY-JUN 2025 [REPEATED] All questions repeated.

Cross-Reference: All Unit 4 Questions Across All 5 Years

Question Topic	MJ-23	ND-23	MJ-24	ND-25	MJ-25
Data Exploration (full 3-tier)	Q3a	-	-	-	Q4a
Univariate/Bivariate/ Multivariate (diff)	-	Q3c	-	-	Q3b
Mean/Median/Mode (computation)	-	Q3a	-	-	-
Data Transformation	Q3b	Q3b	Q3b*	Q3b	Q4b
Data Validation (Incompleteness/Noise/Inconsistency)	Q3c	Q4b	Q3a*	Q3c	Q4c
Contingency Table + Marginal Distribution	-	Q4a	-	Q4a	-
Data Reduction (full detail)	-	-	Q3c	Q3a	Q3a
Sampling, Feature Selection, PCA (reduction)	-	Q4c	-	Q4c	-
Data Discretization / Binning	-	-	Q4b*	Q4b	Q3c
Dirty Data + Reasons	-	-	Q4a	-	-
Binning (step-by-step)	-	-	Q4b	-	-
Bivariate Analysis (dedicated)	-	-	Q4c	-	-

Data Transformation has appeared in all five examination sessions — it is the single most consistently tested topic in Unit 4. Data Discretization has appeared in the last three sessions, indicating it is becoming a high-frequency topic. Contingency Table has appeared twice and is likely to recur.