ASSIGNMENT : HANDLING MISSING DATA IN ETL

Objective

This DPP helps understand:

.Why missing data occurs in ETL pipelines

. How different handling techniques impact analytics

. How to choose the right method instead of blindly deleting data

SECTION A – THEORETICAL QUESTIONS

Q1. What are the most common reasons for missing data in ETL pipelines?

Sol: Missing data in ETL pipelines can occur due to the following major reasons:

1. **Source System Limitations**
   a. The original application may not make certain fields mandatory.
   b. Users can skip optional fields like income, phone number, or address.
2. **Data Extraction Failures**
   a. Network issues, timeouts, or incorrect queries may extract incomplete records.
   b. Large files may get partially read.
3. **Data Integration Problems**
   a. Schema mismatch while merging multiple sources.
   b. Incorrect joins can drop matching values.
4. **User Input Errors**
   a. Typing mistakes, blank submissions, or form abandonment.
5. **File Corruption or Format Issues**
   a. CSV/Excel files may be damaged or contain invalid characters.
6. **Transformation Logic Errors**
   a. Wrong filters, type conversions, or validation rules may convert values to NULL.
7. **Privacy and Compliance Rules**
   a. Sensitive fields may be intentionally masked or removed (GDPR, HIPAA).

Q2. Why is blindly deleting rows with missing values considered a bad practice in ETL?

Sol: Blind deletion is bad because:

- **Loss of valuable information** – Other columns may contain useful data.
- **Reduces dataset size** – Leads to poor model performance and biased analytics.

- **Introduces bias** – Missing data may belong to a specific group (e.g., low-income users).
- **Breaks relationships** – Foreign key or time-series continuity may be lost.
- **Not scalable** – In real ETL, missing values are common and must be handled intelligently.

Q3. Explain the difference between:

- Listwise deletion
- Column deletion
  Also mention one scenario where each is appropriate.

Sol:

1. **Listwise Deletion (Row-Level Removal)**
   - Removes entire row when any selected value is missing.
   - Use only when missing % is very small (<5%).

   Example: Remove rows where Region is blank.

2. **Column Deletion**
   - Remove entire column when > 60–70% values are missing.
   - Use for non-critical attributes.

Appropriate when:

- More than 60–70% of a column is null and the field is not critical (e.g., “fax number”).

Q4. Why is median imputation preferred over mean imputation for skewed data such as income?

Sol: Median imputation is preferred for skewed data like income due to the following reasons:

1. **Mean is affected by outliers**
   a. Income data is usually right-skewed (few people earn very high amounts).
   b. These extreme values pull the mean upward and give unrealistic results.
2. **Median represents typical value better**
   a. Median shows the middle observation and is not influenced by very high or very low incomes.

    b. It reflects the true central tendency of most customers.

3. **Preserves data distribution**
    a. Mean imputation distorts the original distribution and increases bias.
    b. Median keeps the spread of data more natural.
4. **Improves model reliability**
    a. Using mean may overestimate income for low-earning groups.
    b. Median leads to more stable and fair analytics and predictions.

Example:
If incomes are: 20k, 22k, 25k, 27k, **2,00,000**

- Mean = 58,800 (not realistic)
- Median = 25,000 (better representation)

Q5. What is forward fill and in what type of dataset is it most useful?

Sol: **Forward Fill (FFill)** is a missing value handling technique in which the null value is replaced with the **most recent previous valid value** in the dataset.

How it works:

- The last known observation is carried forward until a new value appears.
- No statistical calculation is used—only existing real data is reused.

Most useful in:

1. **Time-Series Data**
    a. Stock prices, temperature records, sensor readings, IoT data.
2. **Sequential Transaction Data**
    a. Bank balance, meter readings, attendance logs.
3. **Datasets where values remain constant for a period**
    a. Customer subscription plan, device status, medical monitoring.

Example:

Time → 10:00 = 35
10:05 = **NULL** → filled with 35
10:10 = 36

Advantages:

- Maintains trend continuity
- Simple and fast
- Suitable when previous value is logically valid

Q6. Why should flagging missing values be done before imputation in an ETL workflow?

Sol: Flagging missing values before performing imputation is an important best practice in ETL for the following reasons:

1. **Preserves Original Information**
    a. Imputation replaces NULL values and hides the fact that data was originally missing.
    b. A flag column (e.g., `Income_Missing = 1/0`) keeps this knowledge intact.
2. **Improves Analytical Accuracy**
    a. The pattern of missingness may itself be meaningful.
    b. Models can learn that "missing income" or "missing address" has business significance.
3. **Supports Better Decision Making**
    a. Different strategies can be applied to records with many missing fields.
    b. Helps in segmentation and risk identification.
4. **Audit and Transparency**
    a. ETL pipelines must be traceable.
    b. Flagging shows which values were original and which were artificially filled.
5. **Prevents Bias**
    a. Imputed values may introduce distortion.
    b. Flag helps analysts treat imputed records separately during reporting.

Q7. Consider a scenario where income is missing for many customers.

How can this missingness itself provide business insights?

Sol: Missing income data should not always be treated as a simple error—it can contain **valuable behavioral information**. The pattern of missingness can reveal important business insights:

1. **Indicator of Customer Sensitivity or Privacy Concerns**
    a. Customers who do not disclose income may be more privacy-conscious.
    b. They may prefer low-risk products and avoid credit or loan services.
2. **Possible Link to Economic Status**
    a. Lower-income groups often skip income fields due to discomfort or fear of rejection.
    b. This segment may require affordable plans or discounts.
3. **Risk Assessment in Financial Services**
    a. In banking, missing income can signal **higher credit risk**.

   b. Such customers can be routed to stricter verification instead of automatic approval.

4. **Product and Marketing Strategy**
   a. Customers with missing income may respond better to prepaid or low-commitment products.
   b. Separate campaigns can be designed for them instead of treating them like high-income users.
5. **Data Quality Monitoring**
   a. A sudden rise in missing income may indicate problems in the data collection form or ETL pipeline.

SECTION B – PRACTICAL QUESTIONS

Customer id	Name	City	Monthly salary	Income	Region
101	Rahul Mehta	Mumbai	12000	65000	WEST
102	Anjali Rao	Bengaluru	NAN	NAN	SOUTH
103	Suresh Lyer	Chennai	15000	72000	SOUTH
104	Neha Singh	Delhi	NAN	NAN	NORTH
105	Amit Verma	Pune	18000	58000	NAN
106	Karan Shah	Ahmedabad	NAN	61000	WEST
107	Pooja Das	Kolkata	14000	NAN	EAST
108	Riya Kapoor	Jaipur	16000	69000	NORTH

Use the given dataset for all questions.

Q8. Listwise Deletion

Remove all rows where Region is missing.

Tasks:

1: Identify affected rows

2:Show the dataset after deletion

3:Mention how many records were lost

Sol :

Step 1: Identify Affected Rows

Listwise deletion removes any record that contains a missing value in the specified column.
In the given dataset, the **Region** column has a missing value for:

- **Customer ID 105 – Amit Verma – Region = NAN**

This row becomes the candidate for deletion.

Step 2: Dataset After Deletion

After removing the row with missing Region, the remaining dataset contains the following customers:

101 – Rahul Mehta – WEST
102 – Anjali Rao – SOUTH
103 – Suresh Iyer – SOUTH
104 – Neha Singh – NORTH
106 – Karan Shah – WEST
107 – Pooja Das – EAST
108 – Riya Kapoor – NORTH

The dataset is now clean with **no missing values in the Region column**.

Step 3: Records Lost

- Total records before deletion = **8**
- Records after deletion = **7**
- **Number of records lost = 1**

Q9. **Imputation**

Handle missing values in Monthly_Sales using:

- **Forward Fill**

Tasks:

1.Apply forward fill

2. Show before vs after values

3. Explain why forward fill is suitable here

Sol: **1. Apply Forward Fill**

Forward fill replaces a missing value with the **previous available value** in the same column.

2. Before vs After Values (Monthly_Salary Column)

Customer id	Name	Monthly Salary Before	Monthly Salary (After ffill)

101	**Rahul Mehta**	**12000**	**12000**
102	**Anjali Rao**	**NAN**	**12000**
103	**Suresh Lyer**	**15000**	**15000**
104	**Neha Singh**	**NAN**	**15000**
105	**Amit Verma**	**18000**	**18000**
106	**Karan Shah**	**NAN**	**18000**
107	**Pooja Das**	**14000**	**14000**
108	**Riya Kapoor**	**16000**	**16000**

Filled Values:

- ID 102 → 12000
- ID 104 → 15000
- ID 106 → 18000

3. Why Forward Fill is Suitable Here

- The data represents customer records in a sequence where nearby customers may belong to similar salary ranges.
- Forward fill uses **real existing values**, not artificial averages.
- It maintains continuity without changing overall distribution.
- Better than mean/median when data is ordered and previous value is a reasonable estimate.

Q10. Flagging Missing Data

Create a flag column for missing Income.

Tasks:

1. Create Income_Missing_Flag (0 = present, 1 = missing)

2.Show updated dataset Count

3.how many customers have missing income

Sol: **Step 1: Create Flag Column**

Rule:**Income_Missing_Flag = 1** → if Income is NAN

- **Income_Missing_Flag = 0** → if Income is present

Step 2: Updated Dataset with Flag

Customer id	**Name**	**Income**	**Income Missing Flag**

101	**Rahul Mehta**	**65000**	**0**
102	**Anjali Rao**	**NAN**	**1**
103	**Suresh Lyer**	**72000**	**0**
104	**Neha Singh**	**NAN**	**1**
105	**Amit Verma**	**58000**	**0**
106	**Karan Shah**	**61000**	**0**
107	**Pooja Das**	**NAN**	**1**
108	**Riya Kapoor**	**69000**	**0**

Step 3: Count of Missing Income

Customers with missing income

- ID **102, 104, 107**

Total missing income = 3 customers