HANDLING MISSING DATA

THEORETICAL QUESTIONS

Question 1: What are the **most common reasons** for missing data in ETL pipelines?

Answer: **Most common reasons for missing data in ETL pipelines:**

- Data not captured at source (user skipped a field)
- System or application errors
- Data corruption during transfer
- Mismatched schemas between source systems
- Manual data entry mistakes
- Late-arriving data

Question 2: Why is **blindly deleting rows with missing values** considered a bad practice in ETL?

Answer: **Blindly deleting rows with missing values is a bad practice:**

- Causes unnecessary data loss
- Can introduce bias in analysis
- Reduces dataset size and reliability
- Missing data itself may carry business meaning

Question 3: Explain the difference between:

- **Listwise deletion**
- **Column deletion**

Also mention **one scenario** where each is appropriate.

Answer: **Difference between Listwise deletion and Column deletion:**

Aspect	**Listwise Deletion**	**Column Deletion**
What it does	Removes entire rows	Removes entire columns
Data loss	High	Very high

When used	Few missing values	Column mostly empty

Scenario:

- **Listwise:** When only 1-2 rows have missing values
- **Column:** When a column has > 70% missing data and is non-critical

Question 4: Why is **median imputation** preferred over mean imputation for skewed data such as income?

Answer: **Median imputation is preferred over mean imputation for skewed data such as income because:**

- Median is resistant to outliers
- Income data is usually right-skewed
- Mean can be misleading due to extreme values

Question 5: What is **forward fill** and in what type of dataset is it most useful?

Answer: **Forward fill replaces missing values using the previous available value.**

Most useful for:

- Time-series data
- Sequential records
- Slowly changing values

Question 6: Why should **flagging missing values** be done before imputation in an ETL workflow?

Answer: **Missing values should be flagged before imputation because:**

- Preserves information about missingness
- Helps detect data quality issues
- Improves model accuracy
- Enables business insights

Question 7: Consider a scenario where income is missing for many customers.

How can this missingness itself provide **business insights?**

Answer: **Missing income can provide business insights:**

- Customers unwilling to disclose income
- Possible high-value customers avoiding sharing details
- Region-specific data collection issues
- Helps target follow-ups or surveys

PRACTICAL QUESTIONS

Question 8: **Listwise Deletion**

Removes all rows where **Region** is missing.

Tasks:

1. Identify affected rows
2. Show the dataset after deletion
3. Mention how many records were lost

Answer: **Affected Row**

- Customer_ID 105 (Region = NaN)

Dataset After Deletion

Customer_ID	Name	City	Monthly_Sales	Income	Region
101	Rahul Mehta	Mumbai	12000	65000	West
102	Anjali Rao	Bengaluru	NaN	NaN	South
103	Suresh Iyer	Chennai	15000	72000	South
104	Neha Singh	Delhi	NaN	NaN	North
106	Karan Shah	Ahmedabad	NaN	61000	West
107	Pooja Das	Kolkata	14000	NaN	East
108	Riya Kapoor	Jaipur	16000	69000	North

Records lost: 1

Question 9: **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

Answer: **Before vs After**

Customer_ID	**Monthly_Sales (Before)**	**Monthly_Sales (After)**
101	12000	12000
102	NaN	12000
103	15000	15000
104	NaN	15000
105	18000	18000
106	NaN	18000
107	14000	14000
108	16000	16000

Why is forward fill suitable

- Data is sequential
- Sales change gradually
- Avoids unrealistic jumps

Question 10: **Flagging Missing Data**

Create a **flag column** for missing income.

Tasks:

1. Create Income _Missing_Flag (0 = present, 1 = missing)
2. Show updated dataset
3. Count how many customers have missing income

Answer: **Income _Missing_Flag**

- **0 = Present**
- **1 = Missing**

Customer_ID	**Income**	**Income_Missing_Flag**
101	65000	0
102	NaN	1
103	72000	0
104	NaN	1
105	58000	0
106	61000	0
107	NaN	1
108	69000	0

Customer with missing income: 3