Name: **Priyanshu Sahu**
Batch: **Data Analytics Nov Live Batch**

Assignment 20: Handling Missing Data in ETL

SECTION A – THEORETICAL QUESTIONS

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

Answer:

Common Reasons for Missing Data in ETL Pipelines:

1. Data entry errors
2. Users skipping optional fields
3. System integration issues
4. Sensor/device failure
5. Data corruption during transfer
6. Different data sources having different formats

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

Answer: Because when we do blindly deleting rows with missing values it leads to:

- Reduces dataset size

- May remove important customers
- Can create biased results
- Reduces accuracy of analytics

Q3.Explain the difference between: Listwise deletion Column deletion Also mention one scenario where each is appropriate

Answer:

Listwise Deletion: Removes entire rows where any value is missing.

Appropriate when missing values are very few and random.

Column Deletion: Removes entire column if most values are missing.

Appropriate when 70–80% data is missing and column is not important.

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

Answer:

We use median imputation preferred over mean imputation for skewed data such as income because:

- Income data is usually right-skewed.
- Mean is affected by extreme values.
- Median is not affected by outliers and gives realistic central value.

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

Answer:

Forward Fill:

Forward fill replaces missing value with the previous available value.

Most useful for time-series, sales data, stock prices, and continuous business records.

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

Answer: **flagging missing values be done before imputation in an ETL workflow because:**

- Missing data itself gives information.

- After filling, we lose that information.
- Flag helps track original missing entries.

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

Answer:

- Customers may belong to a specific segment.
- They may not trust sharing financial details.
- Could indicate low-income group.
- Helps adjust marketing strategy.

SECTION B – PRACTICAL QUESTIONS

Q8. Listwise Deletion Remove all rows where Region is missing.

Answer:

Affected Row: Customer ID 105 (Amit Verma)

Original Records: 8

Records After Deletion: 7

Total Records Lost: 1

Name: **Priyanshu Sahu**
Batch: **Data Analytics Nov Live Batch**

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

Answer:

Before Forward Fill:

101 - 12000

104 - NaN

102 - NaN

105 - 18000

107 - 14000

103 - 15000

106 - NaN

108 - 16000

After Forward Fill:

101	12000
104	12000
102	12000
105	18000
107	14000
103	15000
106	15000
108	16000

Forward fill is suitable because sales data is continuous and previous values can estimate missing ones.

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**

Name: **Priyanshu Sahu**
Batch: **Data Analytics Nov Live Batch**

3. Count how many customers have missing income

Answer:

101 - 65000 - Flag: 0

104 - NaN - Flag: 1

102 - NaN - Flag: 1

105 - 58000 - Flag: 0

107 - NaN - Flag: 1

103 - 72000 - Flag: 0

106 - 61000 - Flag: 0

108 - 69000 - Flag: 0

Total Customers with Missing Income: 3