Handling Missing Data in ETL – Final Submission

Section A: Theoretical Answers (Summarized)

Missing data occurs due to source errors, schema changes, integration failures, corruption, workflow issues, and business rules.

Blind deletion causes data loss, bias, reduced sample size, and business insight loss.

Listwise deletion removes rows; Column deletion removes columns.

Median is preferred over mean for skewed data like income.

Forward fill works best for time-series data.

Flagging preserves missingness information.

Missing income can provide business insights.

Section B: Practical Work

The Excel file contains dataset and operations for:

1. Listwise deletion (Region missing removed)

2. Forward fill on Monthly Sales

3. Income missing flag column

Customer_ID	Name	City	Monthly_Sales	Income	Region
101	Rahul Mehta	Mumbai	12000	65000	West
104	Neha Singh	Delhi			North
102	Anjali Rao	Bengaluru			South
105	Amit Verma	Pune	18000	58000	
107	Pooja Das	Kolkata	14000		East
103	Suresh Iyer	Chennai	15000	72000	South
106	Karan Shah	Ahmedabad		61000	West
108	Riya Kapoor	Jaipur	16000	69000	North