Dataset:

Order_ID	Customer_ID	Sales_Amount	Order_Date
O101	C001	4500	12-01-2024
O102	C002	Null	15-01-2024
O103	C003	3200	2024/01/18
O101	C001	4500	12-01-2024
O104	C004	Three Thousand	20-01-2024
O105	C005	5100	25-01-2024

Q1. Data Understanding

Identify all data quality issues present in the dataset that can cause problems during data loading.

Answer:

Data quality issues in the dataset:

1. Duplicate record exists (Order_ID O101 repeated)
2. Missing value in Sales_Amount (Order_ID O102 = NULL)
3. Invalid data type: “Three Thousand” is text instead of numeric
4. Inconsistent date formats:

DD-MM-YYYY

YYYY/MM/DD

5. Primary key violation due to duplicate Order_ID
6. Dataset is not standardized for loading

Q2. Primary Key Validation

Assume Order_ID is the **Primary Key.**

a) Is the dataset violating the Primary Key rule?

b) Which record(s) cause this violation?

Answer:

a) Yes, the dataset violates the primary key rule.

b) Violating record(s):

Duplicate Order_ID: O101

Q3. Missing Value Analysis

Which column(s) contain **missing values?**

a) List the affected records

b) Explain why loading these records without handling missing values is risky

Answer:

a) Column with missing value:

Sales_Amount

Affected record:

O102

b) Risk of loading missing values:

1. Incorrect sales totals
2. BI dashboards show wrong KPIs
3. Analytical calculations become unreliable
4. Data integrity is compromised

Q4. Data Type Validation

Identify records where **Sales_Amount** violates expected data type rules.
a) Which record(s) will fail numeric validation?
b) What would happen if this dataset is loaded into a SQL table with Sales_Amount as DECIMAL?
Answer:
a) Record failing numeric validation:
O104 → “Three Thousand”
b) If loaded into SQL as DECIMAL:
1. Record will fail
2. Load error or rejection occurs
3. ETL pipeline may stop or skip record
Q5. Date Format Consistency
The Order_Date column has multiple formats.
a) List all date formats present in the dataset
b) Why is this a problem during data loading?
Answer:
a) Date formats present:
12-01-2024 (DD-MM-YYYY)
2024/01/18 (YYYY/MM/DD)
b) Why problematic:
1. Parsing errors during load
2. Wrong sorting and filtering
3. BI tools misinterpret dates
Q6. Load Readiness Decision
Based on the dataset condition:
a) Should this dataset be loaded directly into the database? (Yes/No)
b) Justify your answer with at **least three reasons**
Answer:
a) Should dataset be loaded directly?
No
b) Reasons :
1. Duplicate primary key
2. Missing numeric values
3. Invalid data type
4. Inconsistent date formats
5. Not standardized
Q7. Pre-Load Validation Checklist
List the exact **pre-load validation checks** you would perform on this dataset before loading.
Answer:
1. Primary key uniqueness check
2. Duplicate record detection
3. Missing value validation
4. Numeric data type validation
5. Date format validation

6. Range check for sales values

7. Schema structure validation

8. Referential integrity check

Q8. Cleaning Strategy

Describe the **step-by-step cleaning actions** required to make this dataset load-ready.

Answer:

Step-by-step cleaning:

1. Remove duplicate O101 record
2. Handle NULL value in O102
3. Convert “Three Thousand” → 3000
4. Standardize all dates to YYYY-MM-DD
5. Validate numeric ranges
6. Re-run primary key checks
7. Final audit validation

Q9. Loading Strategy Selection

Assume this dataset represents **daily sales data.**

a) Should a **Full Load** or **Incremental Load** be used?

b) Justify your choice.

Answer:

a) Use: Incremental Load

b) Justification:

1. Dataset represents daily sales
2. Only new records need loading
3. Faster and efficient
4. Reduces processing cost
5. Maintains historical integrity

Q10. BI Impact Scenario

Assume this dataset was loaded **without cleaning** and connected to a BI dashboard.

a) What incorrect results might appear in Total Sales KPI?

b) Which records specifically would cause misleading insights?

c) Why would BI tools not detect these issues automatically?

Answer:

a) Incorrect KPI impact:

1. Duplicate inflates total sales
2. NULL reduces totals
3. Text value breaks calculations
4. Date mismatch causes wrong filtering

b) Misleading records:

O101 (duplicate)

O102 (missing value)

O104 (text amount)

O103 (date inconsistency)

c) Why BI tools won’t detect:

1. BI tools trust input data				
2. No built-in business rule validation				
3. Aggregation happens blindly				
4. Cleaning must occur in ETL laye				