COMPLETE DATASET

Retail Orders & Customer Transactions (Dirty Input)

Business Context

This data comes from a **retail order management system** that integrates:

- Web orders
- Manual entries
- Legacy systems

As expected, the data is **inconsistent, partially corrupt, and unstandardized**.

RAW DATASET (AS RECEIVED)

```
raw_orders = [
    ("ORD001","C001","Ravi"," Delhi ","Laptop","Electronics","45000","202
    ("ORD002","C002","Sneha","Mumbai"," Mobile ","Electronics","32000","0
    ("ORD003","C003","Aman","Bangalore","Laptop","Electronics","55000","2
    ("ORD004","C004","Pooja","Delhi","Tablet"," Electronics ","","2024-01
    ("ORD005","C005","Neha","Chennai","Laptop","Electronics","48000","inv
    ("ORD006","C006","Rahul","Mumbai","Mobile","Electronics",None,"2024-0
    ("ORD007","C007","Kiran","Bangalore","Tablet","Electronics","30000","
    ("ORD008","C008","Amit","Delhi","Laptop","electronics","45000","2024-
    ("ORD009","C009","Priya"," Pune","Mobile","Electronics","28000","09-0
    ("ORD010","C010","Suresh","Mumbai","Laptop","Electronics","55000","20
    ("ORD010","C010","Suresh","Mumbai","Laptop","Electronics","55000","20
    ("ORD011","C011","Meena","Chennai","Tablet","Electronics","31000","20
    ("ORD012","C012","Arjun","Delhi","Mobile","Electronics","27000","2024
    ("ORD013","C013","Nikhil","Bangalore","Laptop","Electronics","60000",
    ("ORD014","C014","Rohit","Mumbai","Mobile","Electronics","invalid_pri
    ("ORD015","C015","Anita","Delhi","Tablet","Electronics","29000","2024
    ("ORD016","C016","Vikas","Chennai","Laptop","Electronics","52000","20
    ("ORD017","C017","Sunita","Mumbai","Mobile","Electronics","33000","20
    ("ORD018","C018","Deepak","Bangalore","Laptop","Electronics","58000",
    ("ORD019","C019","Pallavi","Delhi","Mobile","Electronics","26000","20
    ("ORD020","C020","Manish","Mumbai","Tablet","Electronics","34000","20
]
```

Column Meaning

Column	Description
order_id	Order identifier
customer_id	Customer identifier
customer_name	Customer name
city	Customer city
product	Product name
category	Product category
price	Order price
order_date	Order date
order_status	Status

INTENTIONAL DATA PROBLEMS INCLUDED

String Issues

- Leading / trailing spaces
- Mixed casing (`electronics` , `Electronics`)
- Extra spaces in product and city names

Data Type Issues

- Price as string
- Invalid price values
- Empty strings
- Null values

Date Issues

- Multiple date formats
- Invalid dates

- Mixed separators

Data Quality Issues

- Duplicate records
- Cancelled orders
- Missing prices

CLEANING & TRANSFORMATION TASKS (FOR STUDENTS)

Column Operations

1. Rename all columns to snake_case
2. Add a column `price_with_tax` (18%)
3. Add a column `price_category` (Low / Medium / High)

Data Cleaning

4. Trim and standardize `city`, `product`, `category`
5. Convert price to integer
6. Handle invalid and null prices
7. Normalize all dates into `DateType`
8. Remove duplicate orders
9. Filter only `Completed` orders

Data Transformation

10. Create `order_year`, `order_month`
11. Aggregate total revenue per city
12. Aggregate total revenue per product
13. Identify top 3 cities by revenue
14. Identify products with average price above threshold

File Format Operations

15. Write cleaned data to **Parquet**
16. Read Parquet back and verify schema
17. Write the same data to **ORC**
18. (Optional) Write to **Avro**

Performance & Validation

19. Check number of partitions
20. Repartition before writing
21. Compare file counts between Parquet and ORC
22. Run `explain(True)` on final pipeline
