

Business Context

You are a **Data Engineer** at a retail analytics company.

Sales data is ingested daily from:

- Web applications
- POS systems
- Third-party vendors

Your responsibility is to **clean, standardize, store, and optimize** this data for analytics and downstream systems.

GIVEN DATASET (RAW INPUT)

Create a DataFrame using the following raw data:

```
raw_sales = [  
    ("TXN001", "Delhi ", "Laptop", "Electronics", "45000", "2024-01-05", "Comple  
    ("TXN002", "Mumbai", "Mobile ", "electronics", "32000", "05/01/2024", "Compe  
    ("TXN003", "Bangalore", "Tablet", " Electronics ", "30000", "2024/01/06", "  
    ("TXN004", "Delhi", "Laptop", "Electronics", "", "2024-01-07", "Cancelled")  
    ("TXN005", "Chennai", "Mobile", "Electronics", "invalid", "2024-01-08", "Cc  
    ("TXN006", "Mumbai", "Tablet", "Electronics", None, "2024-01-08", "Complete  
    ("TXN007", "Delhi", "Laptop", "electronics", "45000", "09-01-2024", "Comple  
    ("TXN008", "Bangalore", "Mobile", "Electronics", "28000", "2024-01-09", "Cc  
    ("TXN009", "Mumbai", "Laptop", "Electronics", "55000", "2024-01-10", "Compl  
    ("TXN009", "Mumbai", "Laptop", "Electronics", "55000", "2024-01-10", "Compl  
]
```

Columns

```
transaction_id  
city  
product  
category  
amount
```

transaction_date
status

PART 1 – DATA CLEANING & TRANSFORMATION

1. Trim and standardize all string columns
 2. Convert `category` to uppercase
 3. Convert `amount` to integer
 4. Handle invalid, empty, and null `amount` values
 5. Convert `transaction_date` into `DateTime`
 6. Remove duplicate transactions
 7. Keep only `Completed` transactions
 8. Rename all columns to `snake_case`
-

PART 2 – COLUMN OPERATIONS

9. Add a column `tax_amount` (18% of amount)
 10. Add a column `total_amount` (amount + tax)
 11. Replace city names with standardized values
 12. Rename `transaction_date` to `order_date`
-

PART 3 – ANALYTICS TRANSFORMATIONS

13. Total revenue per city
 14. Total revenue per product
 15. Average order value per city
 16. Top 3 cities by revenue
 17. Identify products with average amount > 40,000
-

PART 4 – PARTITIONS & PERFORMANCE

18. Check current number of partitions
19. Repartition data by `city`
20. Explain why repartitioning is needed

21. Use `explain(True)` and observe the plan

PART 5 – FILE FORMAT STORAGE (HANDS-ON)

Parquet

22. Write cleaned data to **Parquet**

```
data/parquet/sales
```

23. Read Parquet back and validate schema

ORC

24. Write the same data to **ORC**

```
data/orc/sales
```

25. Compare file sizes and number of output files
