Case Study: Data Cleaning & Transformation for Analytics

Background

ABC Retail operates a large network of stores and generates massive daily transaction, customer, and product data. However, raw datasets suffer from duplicates, missing values, inconsistent formats, capitalization errors, and invalid stock entries. These quality issues prevent reliable reporting on revenue, customer engagement, and product performance.

The leadership team expects the Data Engineering function to deliver a trusted, structured dataset that can serve as the foundation for analytics and ML use cases such as customer segmentation and demand forecasting.

Data Sources

- **transactions.csv**: transaction_id, customer_id, product_id, quantity, price, transaction_date, store_id
- **customers.csv**: customer_id, first_name, last_name, email, signup_date, region
- **products.csv**: product_id, product_name, category, cost_price, stock_quantity

Objectives

As a Data Engineer, you are expected to:

1. Build a robust data cleaning and transformation pipeline.
2. Standardize and unify raw data into structured tables.
3. Produce analytics and summary statistics to validate correctness.
4. Deliver modular, reusable logic that can handle future ingestions.

Tasks

Task 1 – Data Cleaning

- Remove duplicates across datasets.
- Handle missing values:
    - Drop transactions with missing quantity or price.
    - Replace missing regions with "Unknown".
    - Convert negative stock values to 0.
- Normalize text (product names, regions) and standardize date formats (YYYY-MM-DD).

Task 2 – Data Transformation

- Create a centralized transaction table with enriched attributes:
  transaction_id, customer_id, customer_name, product_id, product_name, category, quantity, price, transaction_date, store_id, profit
- Compute profit = (price - cost_price) * quantity.
- Aggregate stock data for inventory validation.

Task 3 – Analytics & Insights

- Total transactions per region.
- Total revenue by product category.
- Top 5 products by revenue.
- Customers with missing data (before cleaning).
- *(Optional)* Detect anomalies (e.g., unusually high quantity or price).

Task 4 – SQL Queries

- Transactions per store.
- Products with profit margin <10%.
- Top 3 regions by revenue.

Task 5 – Automation & Documentation

- Implement reusable Python/Pandas scripts.
- Document assumptions, cleaning rules, and transformations.
- Provide summary statistics and validation tables.

Deliverables

- Python/Pandas scripts for cleaning, transformation, and analytics.
- Centralized, cleaned transaction table (CSV or DB export).
- SQL query scripts.
- Documentation: cleaning strategy, transformation logic, and validation summary.

Scoring Rubric

Criteria	Weightage	Evaluation Focus
Data Cleaning & Quality	25%	Correct handling of duplicates, missing values, invalid stock, formatting
Data Transformation Accuracy	25%	Correctness of merged central table, profit computation, schema consistency
Analytics & Summary Statistics	20%	Accuracy of KPIs (transactions, revenue, top products, anomalies)
Code Quality & Modularity	15%	Reusable functions, readability, scalability
Documentation & Communication	15%	Clear explanation of steps, assumptions, and validation evidence