

DATA SCIENCE ENGINEER – ASSESSMENT TEST

PART 1 — Dataset

You will be provided with a CSV file containing sales transactions (5,000 rows).

You must clean, analyze, model, and present insights from the dataset.

PART 2 — Assessment Tasks

1. DATA CLEANING & VALIDATION

- Identify missing, anomalous, or inconsistent values.
- Fix data types.
- Handle duplicates.
- Highlight assumptions and data integrity issues.
- Provide a cleaned CSV.

2. EXPLORATORY DATA ANALYSIS

- Sales trend by hour/day.
- Revenue by category and product.
- Region-wise performance.
- Customer age distribution.
- Online vs retail analysis.
- Provide meaningful visualizations with insights.

3. FEATURE ENGINEERING

Create at least 5 new features:

- Revenue
- Time features (hour, weekday, weekend)
- Age groups
- Product value segmentation
- Any other meaningful transformation

4. MACHINE LEARNING TASK

A. Predict revenue per transaction (Regression)

B. Predict whether a purchase is online vs retail (Classification)

Requirements:

- Train/test split
- Minimum 2 models (RF, XGBoost, Linear/Logistic Regression)
- Performance metrics (MAE, RMSE, Accuracy, F1)
- Explain why one model performs better

5. SQL + SYSTEM DESIGN (15 marks)

SQL:

- Top 3 products by revenue per region
- Total sales per hour
- Categories with avg price > ■5000

System Design:

- How to productionize this ML pipeline?
- How to handle scaling to 5M rows/day?
- How to monitor data drifts and model drifts?
- Time-series forecasting (SMA, ARIMA, LSTM)
- Price/quantity anomaly detection
- Create a prediction API
- Use PySpark for processing