Data Cleaning Report – Amazon Sales Dataset

A. Issues Detected

- Missing values found in critical fields such as Order ID, Date, and Amount.
- Duplicate records detected for some transactions (Order ID + SKU).
- Data type mismatches (e.g., Date stored as text, Amount stored as string).
- Inconsistent formatting in categorical fields such as Category, ship-city, and ship-state.
- Invalid or unrealistic entries (negative quantities and amounts).
- Presence of outliers in Amount values that could distort analysis.

B. Cleaning Actions

- Removed rows with missing values in essential columns; imputed non-critical fields using median or mode.
- Dropped duplicate records to ensure unique transactions.
- Converted data types (Date → datetime, Amount → numeric, Qty → integer).
- Standardized categorical text fields to a consistent format (title case).
- Filtered out invalid values such as negative or zero quantities and amounts.
- Applied IQR method to cap extreme outliers in Amount.

C. New Features Added

- Total Revenue column calculated as Qty × Amount.
- Extracted temporal attributes from Date: Year, Month, Day, Weekday.

D. Assumptions and Considerations

- Orders missing Order ID or Date were excluded as unreliable records.
- Outliers were capped instead of removed to preserve dataset size.
- All transactions are assumed to be recorded in INR currency.

Key Insights

- The dataset is now cleaned, standardized, and analysis-ready.
- Duplicate and invalid entries have been removed.
- Revenue and quantity values are validated, enabling accurate reporting.
- Date-based attributes allow for time-series and seasonal trend analysis.
- The Total Revenue column provides direct visibility into financial performance.