Data Audit Report - transactions.csv

Date: 9 November 2025
Analyst: Nardi

Dataset Overview

- 10,000 rows x 10 columns
- Retail transaction records (October 2025)
- Key fields: Transaction_ID, Customer_ID, Customer_Age, Gender, Country, Transaction_Date, Product_Category, Quantity, Unit_Price, Total_Spend

Data Quality Summary

- Missing values: Found in Customer_Age (178) and Gender (96)
- Duplicates: 25 duplicate rows detected
- Incorrect data types: Transaction_Date stored as string instead of datetime
- Outliers: 8 negative values in Total_Spend, and unusually high Quantity (> 1000 units)
- Inconsistent categories:
- Gender: M, Male, male, F, Female, female
- Country: inconsistent casing (usa, USA, United States)
- Product_Category: Electronics, Electornics, Home Appliance, Home Appliances

Sample Findings

- Transaction_ID = T00987 -> Total_Spend = -42.0
- Transaction_ID = T00456 -> Quantity = 1050
- Gender column includes mixed case variants (male, Male)

Recommended Fixes

- Drop duplicate rows
- Convert Transaction_Date to datetime
- Replace negative Total_Spend values with absolute values or mark as invalid
- Standardize text columns using .str.strip().str.lower() and mapping for consistent category labels
- Impute or drop missing Customer_Age and Gender values depending on business rules

Next Steps

Await manager (Alex) approval to proceed with data cleaning and transformation phase.