Marketing Data Analyst |
Dashboard & BI Developer |
Ex-Head of Marketing & Sales |
Data-Driven Growth | CS Background

PROJECT PORTFOLIO 2026

SQL + Data Cleaning

www.linkedin.com/in/vadayogi | github.com/lgavadayogi

SQL + Data Cleaning: Retail Sales Database Audit

SQL · Python · pandas · SQLite · matplotlib | www.linkedin.com/in/vadayogi | github.com/Igavadayogi

Project Overview

Audited a mid-size U.S.retail company's unreliable sales database before leadership could trust any profitability report.

Conducted a systematic audit of 310 customer records and 2,012 orders, identified 213 data quality issues across 11 categories, then built a 6-step SQL cleaning pipeline to resolve them.

Validated with 7 automated checks before delivering executive business findings.

DATASET: 310 CUSTOMERS · 2,012 ORDERS · 3-YEAR PERIOD (2021-2023)

TOOLS: SQL (SQLITE) · PYTHON (PANDAS, MATPLOTLIB) · JUPYTER NOTEBOOK

APPROACH: AUDIT → CLEAN → VALIDATE → ANALYZE

PIPELINE RESULT:

✓ 2,012 RAW ORDERS → 1,956 VALIDATED RECORDS | 7/7 CHECKS PASSED

Audit Findings

213 issues identified across 11 categories before any cleaning

Issue	Table	Count	Severity
Null ship_date (orders)	orders	81	Medium
Inconsistent segment casing	customers	26	Medium
Inconsistent category casing	orders	20	Medium
Null emails (customers)	customers	18	Medium
Orphaned customer_id (FK)	orders	15	Critical
Duplicate order rows	orders	12	High
Ship date before order date	orders	10	High
Duplicate customer records	customers	10	High
Zero/negative sales	orders	8	High
Null city (customers)	customers	8	Low
Impossible discount > 1.0	orders	5	High

TOTAL: 213 issues | $28,554 in orphaned revenue — unattributable to any customer

SQL + Data Cleaning: Retail Sales Database Audit

Cleaning Pipeline · Validation · BusinessFindings | www.linkedin.com/in/vadayogi | github.com/Igavadayogi

BUSINESS FINDINGS

EAST REGION LEADS

$1.54M revenue · 19.6% profit margin (highest)

Both the highest revenue AND highest margin region.

TECHNOLOGY DOMINATES

$2.81M revenue = 64% of total company revenue

Lowest avg discount (7.1%) = full-price buying.

HOME OFFICE HIGHEST VALUE

$16K revenue per customer vs $13.5K avg

Underserved segment — high value, low volume.

ORPHANED REVENUE RISK

$28,554 unattributable to any customer record

FK violations create blind spots in CRM reporting.

RECOMMENDATION:

Prioritize Home Office segment acquisition — highest individual value, lowest current volume. Investigate and fix orphaned FK records in CRM to restore $28,554 in unattributable revenue.

6-Step SQL Cleaning Pipeline

Raw: 2,012 orders → Clean: 1,956 validated records

Step 1	**Deduplicate customers** 310 → 310 rows (0 removed)
Step 2	**Standardize casing, nulls & dates** Segments · Emails · Cities · Dates
Step 3	**Deduplicate orders** 2,012 → 2,000 rows (12 removed)
Step 4	**Standardize category casing** 10 variants → 5 clean categories
Step 5	**Remove invalid rows** 2,000 → 1,970 rows (30 removed)
Step 6	**Remove orphaned FK records** 1,970 → 1,956 rows (14 removed)

Validation Results — 7/7 Checks Passed

- ✓ No duplicate customer_ids
- ✓ No duplicate order_ids
- ✓ No discount > 1.0
- ✓ No zero/negative sales
- ✓ No orphaned customer_ids
- ✓ All segments valid
- ✓ All categories valid

OUTCOME: $4,342,139 in trustworthy, analyzable revenue across 4 regions

github.com/Igavadayogi/sql-data-cleaning

GET IN TOUCH

PHONE

+62 8113035661

EMAIL

raharjaagungvadayogi@gmail.com

SOCIAL

www.linkedin.com/in/vadayogi

github.com/Igavadayogi

Portfolio Summary

Project 1	**E-Commerce Dashboard** GA4 · Looker Studio	$267K opportunity
Project 2	**Customer Segmentation** Python · K-Means	$1.2M opportunity
Project 3	**Churn Prediction Model** Python · scikit-learn	$231K saved
Project 4	**SQL + Data Cleaning** SQL · Python	213 issues resolved

IGUSTI RAHARJA