Data Quality Assessment Report

Customer Segmentation Project - Online Retail Dataset

Analyst: Gabe Collins
Date: 10/26/25
Dataset: UCI Online Retail Dataset (Dec 2010 - Dec 2011)
Source: https://archive.ics.uci.edu/dataset/352/online+retail

---

Executive Summary

This report documents data quality issues identified in the Online Retail dataset and the remediation steps taken to ensure reliable analysis.

Key findings:

- **Total records:** 541,909 transactions
- **Records removed:** 153,144 (28.3%)
- **Final clean dataset:** 388,765 transactions (71.7%)
- **Impact:** While cleaning removed 28% of transactions and 31% of revenue, it only impacted 2% of customers. This indicates the removed data consisted primarily of:
    - Cancelled/returned orders
    - Wholesale/bulk orders (extreme outliers)
    - Data quality issues (missing customer IDs, duplicates, others listed below)

The 3.9% decrease in average transaction value confirms that cleaning successfully isolated typical retail customer behavior by removing wholesale/B2B transactions and anomalies.

---

1. Baseline/Missing Data Analysis

1.1 Missing Customer IDs

Issue Identified:

```
SELECT
  COUNT(*) as total_records,
  COUNT(*) - COUNT(customer_id) as missing_customer_id,
  ROUND((COUNT(*) - COUNT(customer_id))::NUMERIC / COUNT(*) * 100, 2) as pct_missing
```

```
FROM retail.online_retail_raw;
```

Results:

- **Missing Customer IDs:** 135,080 records (24.9%)
- **Impact:** Cannot perform customer segmentation without customer identifier

Decision: REMOVED records with missing ids

- **Rationale:** Customer-level analysis requires customer_id; these records cannot be used for segmentation
- **Business Impact:** Minimal - these may be guest checkouts or data entry errors
- **Alternative Considered:** Could analyze as "Anonymous" segment, but rejected due to limited value

Code Applied:

```
WHERE customer_id IS NOT NULL
```

1.2 Duplicate Transactions

Issue Identified:

```
SELECT
    invoice_no,
    stock_code,
    customer_id,
    COUNT(*) as duplicate_count
FROM retail.online_retail_raw
WHERE customer_id IS NOT NULL
GROUP BY invoice_no, stock_code, customer_id
HAVING COUNT(*) > 1;
```

Result: 5,225 duplicates found across 1,903 invoices

Decision: REMOVED duplicates (pending validation from the ecomm team or further investigation into how the ecomm system collects data)

- **Investigation:**
    - Upon detailed examination, these records contain identical information across every column

    - There are records outside of these 5,225 that contain identical information across all columns EXCEPT the quantity column
        - It is possible that customers added the exact same product to their cart multiple times pulling into the invoice as multiple line items (will check with appropriate team)
- **Rationale:**
    - Exact duplicates inflate transaction counts and revenue without adding value.
    - Partial duplicates represent legitimate business scenarios (e.g., customer ordered same item in different quantities/colors/sizes within same invoice).
- **Impact:**
    - Records removed: 5,225 records - a small portion (1%) of the dataset
    - Revenue impact: £21,546.39 (0.26% of remaining revenue)
- **Next Steps / Follow-up:**
    - Reach out to appropriate team for better understanding of system

---

2. Data Anomalies

2.1 Cancelled/Returned Orders

Issue Identified:

```
SELECT
   COUNT(*) as cancelled_orders,
   ROUND(COUNT(*)::NUMERIC / (SELECT COUNT(*) FROM retail.online_retail_raw
                    WHERE customer_id IS NOT NULL) * 100, 2) as pct_cancelled
FROM retail.online_retail_raw_deduped
WHERE customer_id IS NOT NULL
  AND invoice_no LIKE 'C%';
```

Results:

- **Cancelled Orders:** 8,905 records (2.3% of remaining records)
- **Dataset Info:** Invoice numbers starting with 'C' indicate cancellations/returns

Decision: REMOVED records of cancelled/returned orders

- **Rationale:** Cancellations represent negative transactions that distort customer value
- **Business Context:** These are refunds/returns, not actual purchases
- **Impact on Metrics:**

    - Including them would understate customer value
    - Could create negative transaction amounts
- **Alternative Approach:** Could analyze separately for return rate analysis

Code Applied:

```
WHERE invoice_no NOT LIKE 'C%'
```

2.2 Negative Quantities

No Issue Identified:

```
SELECT
   COUNT(*) FILTER (WHERE quantity < 0) as negative_quantity,
   MIN(quantity) as min_quantity,
   ROUND(COUNT(*) FILTER (WHERE quantity < 0)::NUMERIC / COUNT(*) * 100, 2) as pct_negative
FROM retail.online_retail_raw_deduped
WHERE customer_id IS NOT NULL
 AND invoice_no NOT LIKE 'C%';
```

Results:

- **Negative Quantities:** 0 records (0%)

Decision: Add code for future data integrity

- **Rationale for checking:** Negative quantities likely represent returns not marked with 'C' prefix
- **Data Integrity:** Consistent with positive revenue analysis

Code Applied:

```
WHERE quantity > 0
```

2.3 Negative or Zero Unit Prices

Issue Identified:

```
SELECT
```

```
    COUNT(*) FILTER (WHERE unit_price < 0) as invalid_price,
    MIN(unit_price) as min_price,
    MAX(unit_price) as max_price,
    COUNT(*) FILTER (WHERE unit_price = 0) as zero_price
FROM retail.online_retail_raw_deduped
WHERE customer_id IS NOT NULL
  AND invoice_no NOT LIKE 'C%'
  AND quantity > 0;
```

Results:

- **Zero Prices:** 44 records
- **Issue:** Zero prices may indicate promotional items, data errors, or manual adjustments

Decision: REMOVED zero price records

- **Rationale:** Cannot calculate accurate revenue or customer value with £0 prices
- **Business Logic:** Even promotional items have a cost; £0 suggests data error
- **Impact:** Small volume, minimal impact on overall analysis

Code Applied:

```
WHERE unit_price > 0
```

---

3. Outlier Analysis

3.1 Extreme Transaction Values

Issue Identified:

```
WITH transaction_values AS (
    SELECT
        quantity * unit_price as total_price
    FROM retail.online_retail_raw_deduped
    WHERE customer_id IS NOT NULL
      AND invoice_no NOT LIKE 'C%'
      AND quantity > 0
      AND unit_price > 0
)
```

```
SELECT
  PERCENTILE_CONT(0.50) WITHIN GROUP (ORDER BY total_price) as median,
  PERCENTILE_CONT(0.95) WITHIN GROUP (ORDER BY total_price) as p95,
  PERCENTILE_CONT(0.99) WITHIN GROUP (ORDER BY total_price) as p99,
  MAX(total_price) as max_value
FROM transaction_values;
```

Results:

- **P95 (95th percentile):** £67.50
- **P99 (99th percentile):** £203.52
- **Maximum value:** £168,469.60
- **Top 1% range:** £204 - £168,470

Analysis:

Percentile	Value	Interpretation
P50 (Median)	£11.80	Typical transaction
P95	£67.50	Large but reasonable
P99	£203.52	Very large order
Max	£168,469.60	Extreme outlier (~831x larger than P99)

Decision: REMOVED TOP 1%

- **Rationale:**
  - Top 1% contains extreme bulk/wholesale orders
  - These orders are 100-800x larger than typical retail transactions
  - May represent B2B sales rather than individual customers
  - Would distort "typical customer" segmentation
- **Business Context:** Dataset description mentions wholesale customers; these are likely B2B
- **Impact:**
  - Removed 3,890 transactions (1% of remaining)
  - Retained 99% of transaction volume
  - Analysis now focuses on retail customer behavior
- **Alternative:** Could create separate B2B segment (for more advanced analysis)

Code Applied:

```
WHERE (quantity * unit_price) <= (
  SELECT PERCENTILE_CONT(0.99) WITHIN GROUP (ORDER BY quantity * unit_price)
  FROM retail.online_retail_raw
```

```
)
```

---

4. Data Consistency Checks

4.1 Date Range Validation

Check Performed:

```
SELECT
    MIN(invoice_date) as earliest_transaction,
    MAX(invoice_date) as latest_transaction,
    EXTRACT(DAY FROM (MAX(invoice_date) - MIN(invoice_date))) as days_covered
FROM retail.online_retail_raw;
```

Results:

- **Date Range:** December 1, 2010 to December 9, 2011
- **Coverage:** 373 days (~12.5 months)
- **Status:** Valid - matches dataset description

4.2 Customer ID Format Validation

Check Performed:

```
SELECT
    MIN(customer_id) as min_id,
    MAX(customer_id) as max_id,
    COUNT(DISTINCT customer_id) as unique_customers
FROM retail.online_retail_raw
WHERE customer_id IS NOT NULL;
```

Results:

- **Range:** 12346 to 18287
- **Unique Customers:** 4,372
- **Format:** All numeric, 5-digit IDs - consistent

---

5. Cleaning Summary

Before & After Comparison

Metric	Raw Data	Clean Data	Change
Total Records	541,909	388,765	-153,144 (-28.3%)
Unique Customers	4,372	4,290	-82 (-1.9%)
Date Range	Dec 2010 - Dec 2011	Dec 2010 - Dec 2011	No change
Total Revenue	£9,747,747	£6,722,262	-£3,025,485 (-31.0%)
Avg Transaction	£17.99	£17.29	-£0.70 (-3.9%)

Key Observations:

- Date range fully preserved
- Minimal customer loss:
  - Retained 98% of unique customers (removed B2B & large transactions)
- Significant transaction removal (28.3%)
  - 135,080 of the 153,223 removed records were missing customer IDs
- Significant revenue removal (31%)
  - £2,167,599 removed with top 1% outliers
  - £1,447,682 removed with missing customer IDs
  - £611,342 added from cancelled orders
- Average transaction value decreased slightly (confirming I removed high-value outliers/errors, not merely typical data)
- Dataset now suitable for reliable customer segmentation

---

6. Data Quality Rules Applied

SQL Cleaning Filter (Final)

```
SELECT
    invoice_no,
    stock_code,
    description,
    quantity,
    invoice_date,
    unit_price,
```

```
    customer_id,
    country,
    (quantity * unit_price) as total_price,
    DATE_TRUNC('month', invoice_date) as invoice_month,
    EXTRACT(DOW FROM invoice_date) as day_of_week,
    EXTRACT(HOUR FROM invoice_date) as hour_of_day
FROM retail.online_retail_raw_deduped       -- Remove duplicates
WHERE
    customer_id IS NOT NULL                 -- Remove missing customers
    AND invoice_no NOT LIKE 'C%'            -- Remove cancellations
    AND quantity > 0                        -- Remove returns/negatives
    AND unit_price > 0                      -- Remove zero/negative prices
    AND (quantity * unit_price) <= (        -- Remove top 1% outliers
        SELECT PERCENTILE_CONT(0.99) WITHIN GROUP (ORDER BY quantity * unit_price)
        FROM retail.online_retail_raw_deduped
    );
```

---

7. Recommendations for Future Data Collection

Based on this quality assessment, I recommend:

Critical Priority (Immediate Implementation)

1. **Implement Customer ID Validation**
    - Issue:
        - 135,080 transactions (24.9%) missing customer_id
        - Prevents customer-level analysis for 1/4 of all transactions
    - Proposed Solution:
        - Make customer_id a required field at point of sale/checkout
        - Implement system-level validation to prevent transaction submission without customer_id
        - For guest checkouts, auto-generate unique guest IDs
    - Expected Impact:
        - Reduce missing customer_id from 24.9% to <1%
        - Enable analysis of 100% of transactions
        - Improve customer lifetime value accuracy

        - Better identify returning customers
    - Implementation Cost:
        - Low (configuration change in existing systems)

2. **Implement Transaction Type Classification**
    - Issue:
        - Cannot distinguish B2C (retail) from B2B (wholesale) transactions
        - Mixed customer types distort segmentation analysis
    - Proposed Solution:
        - Add customer_type field: "Retail" vs. "Wholesale"
        - Add transaction_type field: "Sale", "Return", "Exchange"
        - Implement at customer registration or first purchase
        - Auto-flag transactions >£500 as "Potential Wholesale" for review
    - Expected Impact:
        - Enable separate analysis of B2C vs B2B customers
        - Improve accuracy of retail customer segmentation
        - Better target marketing campaigns by customer type
        - Easier identification of high-value wholesale accounts
    - Implementation Cost: Medium (requires database schema change and UI updates)

High Priority (Short-Term Implementation)

3. **Standardize Cancellation Process**
    - Current Issue:
        - 8,905 cancelled orders identified by 'C' prefix in invoice_no
        - Some negative quantities without 'C' prefix (inconsistent marking)
        - No return reason code or timestamp
    - Proposed Solution:
        - Create dedicated order_status field: "Completed", "Cancelled", "Returned", "Partial Return"
        - Add return_reason field with standardized codes:
            - "Customer Request"
            - "Defective Product"
            - "Wrong Item Shipped"
            - "Changed Mind"
            - "Other"
        - Record cancellation_date and return_date separately from original invoice_date
        - Maintain original transaction record, add linked cancellation record
    - Expected Impact:

        - Clear separation of sales vs. returns in reporting
        - Analyze return patterns by reason
        - Calculate accurate return rates by product/customer
        - Improve inventory management based on return reasons
    - Implementation Cost: Medium (database schema + workflow changes)

4. **Implement Data Validation Rules**
    - Current Issue:
        - Negative quantities (10,624 records)
        - Zero or negative prices (1,338 records)
        - Extreme outliers (max transaction 831x larger than P99)
    - Proposed Solution:
        - Price Validation:
            - Set minimum price threshold (e.g., >£0.01) at product catalog level
            - Flag £0 prices as "Promotional" or "Sample" with special code
            - Prevent negative prices in system
            - Alert on prices >3 standard deviations from product's historical average
        - Quantity Validation:
            - Prevent negative quantities (use return process instead)
            - Set maximum quantity limits per product (e.g., 1,000 units)
            - Alert on quantities >99th percentile for review
        - Transaction Amount Validation:
            - Flag transactions >£5,000 as "Requires Review"
            - Separate approval workflow for wholesale orders
    - Expected Impact:
        - Eliminate data entry errors at source
        - Faster identification of fraudulent transactions
        - Cleaner data for real-time analytics
    - Implementation Cost: Low-Medium (validation rules in existing system)

Medium Priority (Long-Term Improvements)

5. **Add Customer Demographics**
    - Current Issue:
        - Limited customer information (only ID and country)
        - Cannot segment by age, gender, or other demographics
        - No acquisition channel tracking
    - Proposed Solution:
        - Collect during account registration (optional fields):

        - Age range (dropdown: 18-24, 25-34, 35-44, 45-54, 55-64, 65+)
        - Gender (optional)
        - Customer since date (auto-populated)
    - Track acquisition source:
        - "Organic Search"
        - "Paid Ad"
        - "Email Campaign"
        - "Social Media"
        - "Direct"
        - "Referral"
    - Add email_opt_in flag for marketing consent
- Expected Impact:
    - Richer customer segmentation (beyond RFM)
    - Demographic-based marketing campaigns
    - Better understanding of customer acquisition ROI
    - Personalized product recommendations
- Implementation Cost: Medium (requires customer-facing changes)

---

8. Impact on Analysis

Reliability Assessment

Analysis Type	Data Quality Impact	Confidence Level
RFM Segmentation	High - Clean customer-level data	Very High
Customer Lifetime Value	High - Accurate revenue totals	Very High
Cohort Analysis	High - Complete date coverage	Very High
Product Analysis	Medium - Some missing descriptions	Medium-High
Time Series	High - Consistent date range	Very High

Key Assumptions

1. **Missing Customer IDs:** Assumed to be guest checkouts or data entry errors; not recoverable
2. **Outliers:** Top 1% assumed to be B2B/wholesale rather than typical retail customers
3. **Cancellations:** All 'C' prefixed invoices are cancellations/returns; some non-prefixed negatives also caught
4. **Data Completeness:** December 2011 data only through Dec 9 (partial month)

---

9. Files Generated

- data_quality_checks.sql - All validation queries
- data_cleaning.sql - Cleaning transformation
- online_retail_clean.csv - Cleaned dataset
- cleaning_summary.xlsx - Before/after metrics

---

10. Sign-Off

Data Quality Status: ✅ APPROVED FOR ANALYSIS

The dataset has been successfully cleaned and validated. The remaining 388,765 transactions across 4,290 customers provide a reliable foundation for customer segmentation analysis.

Analyst: Gabe Collins
Date: 10/27/25
Reviewed By: Gabe Collins

Appendix A: SQL Validation Queries

All queries used in this assessment are available in:

- sql/data_quality_checks.sql
- cleaning_summary.xlsx