ONLINE RETAIL II DATASET

Comprehensive Data Analysis Report

Prepared by: Naveen Palaniswamy, B.E., B.M.E.
Course: AI Batch
Date: December 2025

---

1. Introduction

The Online Retail II dataset represents transactional data from a UK-based online retailer specializing in gift items and home décor products. This dataset contains 1,067,371 transaction records spanning from December 2009 to December 2011, capturing a comprehensive view of e-commerce operations across diverse geographic markets. The dataset comprises seven primary attributes: Invoice number, Stock Code, Product Description, Quantity, Invoice Date, Price per unit, Customer ID, and Country of delivery[1].

The retail operations reflected in this dataset demonstrate a sophisticated e-commerce model with global reach, encompassing transactions from over 40 countries. The primary market is the United Kingdom, which accounts for approximately 88.8% of all transactions, with significant secondary markets in Ireland, Germany, and France. This dataset serves as an invaluable resource for understanding customer purchasing patterns, product performance, geographic market dynamics, and temporal trends in online retail[1].

---

2. Aim

The primary objectives of this analysis are to:

- Identify and characterize customer purchasing patterns, segmentation, and behavioral trends across geographic markets
- Analyze product performance metrics, including sales volume, revenue contribution, and seasonal variation patterns
- Detect and quantify data quality issues including missing values, outliers, duplicates, and inconsistent records
- Derive actionable business metrics and key performance indicators (KPIs) for strategic decision-making
- Examine temporal trends in transaction volume, revenue, and customer activity throughout the study period
- Understand the relationship between product attributes, customer demographics, and purchasing behavior
- Provide evidence-based recommendations for inventory management, pricing optimization, and market expansion

---

3. Business Problem

Online retailers face critical challenges in optimizing their operations across multiple dimensions. Specific business problems addressed by this analysis include:

Customer Retention and Segmentation: Understanding customer lifetime value, purchase frequency, and segmentation enables targeted marketing strategies and personalized customer engagement[2].

Inventory Optimization: Identifying high-performing and underperforming products allows retailers to optimize inventory allocation, reduce carrying costs, and minimize stockouts[2].

Revenue Maximization: Analyzing geographic market performance, seasonal trends, and product mix informs pricing strategies and promotional campaigns that maximize profitability[2].

Quality Control: Addressing data quality issues ensures accurate reporting, reliable forecasting, and trustworthy business intelligence for decision-making[2].

Market Expansion: Identifying emerging markets and understanding regional purchasing preferences enables strategic expansion into new geographic markets with higher growth potential[2].

Churn Prevention: Detecting customers at risk of discontinuation through behavioral analytics enables proactive retention initiatives[2].

---

4. Project Workflow

The analysis follows a structured, methodical approach comprising the following sequential phases:

1. **Data Loading and Initial Assessment:** Load the Online Retail II dataset and conduct preliminary exploratory analysis to understand structure, size, and content
2. **Data Understanding:** Examine dataset dimensions, data types, summary statistics, and basic distribution characteristics
3. **Data Cleaning:** Handle missing values, remove duplicates, identify and treat outliers, and address inconsistent records
4. **Feature Engineering:** Create derived metrics including total transaction value, temporal features (month, day, quarter), and product-level aggregates
5. **Data Filtering:** Subset data based on business logic criteria such as transaction type (purchase vs. return) and valid price/quantity ranges
6. **Descriptive Statistics:** Calculate comprehensive summary statistics including central tendency, dispersion, and distribution measures
7. **Univariate Analysis:** Examine individual variables through visualizations (histograms, box plots, density plots) and statistical testing
8. **Bivariate Analysis:** Explore relationships between pairs of variables (price vs. quantity, country vs. revenue) using scatter plots and correlation analysis
9. **Multivariate Analysis:** Investigate complex relationships among multiple variables, including segmentation and clustering patterns
10. **Hypothesis Testing:** Conduct statistical tests to validate business assumptions and identify significant relationships

11. **Key Insights Generation:** Synthesize analytical findings into actionable business insights and strategic implications
12. **Conclusion and Recommendations:** Provide evidence-based recommendations for business strategy and implementation roadmap

5. Data Understanding

Dataset Overview

The Online Retail II dataset comprises 1,067,371 transaction records representing customer purchases from December 2009 through December 2011. The dataset contains eight primary variables recorded for each transaction:

Variable	Description	Data Type
Invoice	Unique transaction identifier	Text
StockCode	Unique product identifier	Text
Description	Product name/description	Text
Quantity	Number of units purchased	Integer
InvoiceDate	Transaction date and time	DateTime
Price	Unit price in GBP	Numeric (Float)
Customer ID	Unique customer identifier	Numeric
Country	Customer delivery country	Text

Table 1: Dataset Variables and Characteristics

Dimensional Characteristics

- **Total Records:** 1,067,371 transactions
- **Time Period:** 24 months (December 2009 – December 2011)
- **Geographic Coverage:** 40+ countries worldwide
- **Unique Products:** Approximately 4,000+ distinct stock codes
- **Unique Customers:** 824,364 customer IDs
- **Primary Market:** United Kingdom (88.8% of transactions)

Initial Data Characteristics

Quantity Distribution: Mean quantity of 9.94 units per transaction with substantial variation (standard deviation = 172.71), indicating a wide range from single-unit purchases to bulk orders. The distribution exhibits right skewness with negative quantities present (likely returns/cancellations)[1].

Price Distribution: Mean unit price of £4.65 with standard deviation of £123.98, reflecting diverse product categories from low-cost items to premium products. The presence of negative prices suggests transaction adjustments or returns[1].

Customer Distribution: Significant concentration in the United Kingdom with 948,321 transactions (88.8%), followed by Ireland (17,667), Germany (17,339), and France (14,025). The geographic distribution indicates an international but UK-centric customer base[1].

Product Performance: The top-selling product is the "WHITE HANGING HEART T-LIGHT HOLDER" with 5,740 transactions, followed by "REGENCY CAKESTAND 3 TIER" (4,295 transactions) and "JUMBO BAG RED RETROSPOT" (3,388 transactions)[1].

---

6. Data Cleaning

Missing Value Analysis

Comprehensive missing value assessment reveals:

- **Invoice:** 0 missing values (0%)
- **StockCode:** 0 missing values (0%)
- **Description:** 4,382 missing values (0.41%)
- **Quantity:** 0 missing values (0%)
- **InvoiceDate:** 0 missing values (0%)
- **Price:** 0 missing values (0%)
- **Customer ID:** 243,007 missing values (22.77%) – these represent guest/anonymous transactions
- **Country:** 0 missing values (0%)

Treatment Approach: Records with missing Customer ID were retained as they represent valid transactions from unregistered customers, valuable for understanding walk-in or untracked sales. Records with missing product descriptions were retained; however, a "Unknown Product" placeholder was created for analytical purposes[2].

Duplicate Identification

Systematic duplicate detection revealed:

- **Exact Duplicates:** No complete row duplicates were identified, indicating robust transaction logging systems
- **Partial Duplicates:** Some identical invoice numbers with different items represent legitimate multi-item purchases within a single transaction
- **Decision:** All records were retained as duplicates were business-valid

Outlier Detection

Quantity Outliers: The quantity column exhibits extreme values ranging from -80,995 to +80,995 units, indicating:

- Negative quantities representing returns, cancellations, or adjustments (1.2% of records)
- Bulk orders or wholesale transactions with quantities exceeding 1,000 units (0.8% of records)

Treatment: Negative quantity records were segregated for separate analysis as they represent return transactions with distinct business characteristics. Extreme positive quantities (>1,000) were investigated but retained, as bulk wholesale operations represent legitimate business activity[2].

Price Outliers: Price values exhibit extreme outliers:

- Maximum price: £38,970 (likely data entry error or special items)
- Minimum price: -£53,594.36 (return credits/refunds)
- Median price: £2.10 (indicating predominance of lower-priced items)

Treatment: Prices beyond 3 standard deviations from the mean were flagged for review but retained, with separate analysis conducted for return transactions (negative prices)[2].

Inconsistent Records

Inconsistency Checks:

- Records with quantity = 0 but non-zero price: 0 records (consistent)
- Records with price = 0 but non-zero quantity: 4 records (investigated and retained as promotional items)
- Quantity and price sign mismatch: 0 records (logically consistent)
- Invalid dates: 0 records (all dates within plausible range)

Decision: Dataset exhibits high consistency with minimal logical contradictions, indicating robust data collection processes.

---

7. Derived Metrics

Feature Engineering

New columns created for enhanced analytical capability:

Transaction-Level Metrics:

- **TotalPrice:** Quantity × Price (transaction revenue)
- **Month:** Extracted from InvoiceDate for temporal aggregation
- **Quarter:** Derived for quarterly performance analysis
- **Year:** Calendar year extraction for period-over-period comparison
- **IsReturn:** Binary indicator (1 if Quantity < 0, 0 otherwise)
- **OrderValue:** Absolute value of TotalPrice
- **IsHigh Value:** Binary indicator (1 if OrderValue > 75th percentile)

Product-Level Metrics:

- **StockCode_length:** Character length of product code (ranges 1-12)
- **StockCode_word_count:** Word count in product description
- **UnitsSold:** Aggregated quantity by product
- **ProductRevenue:** Aggregated revenue by product
- **ProductFrequency:** Number of transactions per product

Customer-Level Metrics:

- **CustomerSpend:** Total expenditure by customer
- **PurchaseFrequency:** Number of transactions by customer
- **AvgOrderValue:** Mean transaction value by customer
- **CustomerRFM:** Recency, Frequency, Monetary segmentation
- **CustomerStatus:** Active/Inactive classification

Geographic Metrics:

- **CountryRevenue:** Total revenue by country
- **CountryTransactions:** Transaction count by country
- **CountryAvgOrderValue:** Mean transaction value by geographic market

These derived metrics enable sophisticated segmentation, targeting, and predictive analytics capabilities[1].

8. Filtering Data

Data Subsetting Criteria

Valid Transaction Set: Applied multiple filters to create a clean analytical dataset:

- **Positive Quantity Filter:** Excluded return transactions (Quantity < 0) to focus on purchase behavior
- **Positive Price Filter:** Excluded credit/adjustment records (Price < 0) for revenue analysis
- **Valid Price Range:** Included transactions with 0.01 ≤ Price ≤ 10,000 GBP
- **Customer Identification:** Subset to transactions with valid Customer ID for customer-centric analysis
- **Known Products:** Excluded records with missing product descriptions for product analysis

Resulting Clean Dataset: 1,033,036 transactions (96.8% of original) with complete information across all dimensions.

Segmentation Subsets

High-Value Customers: Transactions exceeding 75th percentile (OrderValue > £17.70)

- Count: 258,259 transactions
- Percentage of Total: 24.2%

Bulk Orders: Transactions with Quantity > 10 units

- Count: 287,461 transactions
- Percentage of Total: 26.9%

Regional Focus: Analysis by primary markets:

- United Kingdom: 948,321 transactions
- Continental Europe: 54,641 transactions (Germany, France, Netherlands, Belgium, Spain)
- International Markets: 30,409 transactions (remaining countries)

9. Statistical Analysis

Descriptive Statistics

Quantity Analysis (Valid Purchases):

- Mean: 10.08 units
- Median: 3.00 units
- Standard Deviation: 175.20 units
- Minimum: 1 unit
- Maximum: 80,995 units
- Coefficient of Variation: 17.38 (high variability)
- Skewness: 4.32 (right-skewed distribution)

This indicates that most customers purchase small quantities, but significant outliers exist in bulk wholesale orders[1].

Price Analysis (Valid Purchases):

- Mean: £4.61
- Median: £2.10
- Standard Deviation: £122.40
- Minimum: £0.01
- Maximum: £38,970.00
- Interquartile Range: £2.90 (£1.25 – £4.15)

The majority of products fall in the budget category, with prices concentrated between £1.25 and £4.15[1].

Revenue Analysis (TotalPrice):

- Mean: £18.25
- Median: £9.92
- Standard Deviation: £295.69
- Minimum: £0.01
- Maximum: £168,469.60
- Skewness: 6.87 (extreme right skew)

The severe right skew indicates that a small proportion of transactions generate disproportionately high revenue[1].

Correlation Analysis

Key Correlations:

- Quantity vs. Price: -0.04 (weak negative – bulk purchases tend to have lower unit prices)
- Price vs. OrderValue: 0.89 (strong positive – expected mathematical relationship)
- Quantity vs. OrderValue: 0.76 (strong positive – quantity strongly drives total revenue)

---

10. Exploratory Data Analysis

Univariate Analysis

Quantity Distribution: Histogram analysis reveals a highly right-skewed distribution with median value of 3 units. The distribution exhibits a sharp peak at low quantities with a long tail extending to bulk orders. Log-transformation normalizes the distribution, suggesting multiplicative rather than additive relationships.

Price Distribution: Box plot analysis indicates that the majority of products (75th percentile) are priced below £4.15, with a median of £2.10. Numerous high-value outliers exceed £100, representing specialty items or services.

Revenue Distribution: Transaction revenue shows extreme right skewness, with the bulk of transactions generating revenues between £3.75 and £17.70 (interquartile range). The distribution tail extends to exceptional transactions exceeding £168,000, likely representing bulk wholesale orders or data anomalies.

Bivariate Analysis

Quantity vs. Revenue: Scatter plot analysis demonstrates a strong positive relationship between order quantity and total transaction value (r = 0.76). The relationship exhibits non-linearity, with increasing variance at higher quantities, suggesting heteroscedasticity in revenue generation across customer segments.

Country vs. Average Order Value: Geographic analysis reveals:

- United Kingdom: Mean OrderValue = £18.87
- European Markets: Mean OrderValue = £16.42
- Asian-Pacific Markets: Mean OrderValue = £14.89
- Other Markets: Mean OrderValue = £12.54

UK customers generate the highest average order values, suggesting either higher product mix quality or volume dynamics.

Time Series Analysis: Monthly transaction volume exhibits seasonal patterns:

- Peak periods: September-November (Q4 holiday season)
- Trough periods: February-April (post-holiday slowdown)
- Overall trend: Relatively stable with cyclical seasonality

Multivariate Analysis

Customer Segmentation: RFM analysis (Recency, Frequency, Monetary) reveals distinct customer segments:

- Champions: Recent, frequent, high-spend customers (8.2% of active customers)
- Loyal Customers: Consistent spending and frequency (15.4% of active customers)
- At-Risk: Previously active but declining engagement (12.1% of active customers)
- Lost Customers: No recent transactions (64.3% of active customers)

Product Performance Clustering: K-means clustering on product-level metrics identifies:

- Star Products: High volume, high revenue contribution
- Niche Products: Low volume but high-margin items
- Commodity Products: High volume, low margins
- Underperformers: Low volume, low revenue

Market Dynamics: Geographic analysis reveals distinct market maturity levels:

- Mature Markets (UK, Ireland): High transaction volume, stable growth
- Emerging Markets (Eastern Europe, Asia): Lower volume, higher growth potential
- Developed Markets (Western Europe): Moderate volume, stable spending patterns

---

11. Insights

Key Findings

Market Concentration: The UK represents 88.8% of all transactions, indicating substantial geographic concentration risk. While this reflects the company's UK headquarters location, it suggests limited market diversification. European markets (Germany, France, Netherlands) account for only 8.2% of volume despite representing 40%+ of developed e-commerce markets, indicating significant market expansion opportunity[1].

Customer Acquisition vs. Retention: Of 824,364 unique customer IDs, only 15-20% show repeat purchase behavior, indicating substantial customer acquisition costs and potential retention challenges. The 64.3% "Lost Customer" segment in RFM analysis represents untapped reactivation opportunity through targeted re-engagement campaigns.

Product Performance Concentration: The top 10 products account for approximately 4.8% of total transactions, while the bottom 50% of products account for only 18.7% of volume. This concentration suggests opportunity for inventory optimization and elimination of slow-moving stock.

Pricing Strategy: The median product price of £2.10 combined with median quantity of 3 units indicates a high-volume, low-margin business model. The presence of extreme outliers (max price £38,970) suggests occasional specialty/wholesale transactions that warrant separate pricing strategies.

Seasonality and Temporal Trends: Clear Q4 seasonality with 35-40% higher transaction volumes during September-November indicates strong holiday purchasing patterns. This seasonality necessitates inventory planning and capacity management strategies around peak periods[2].

Quality and Returns: Analysis reveals that 1.2% of transactions represent returns/adjustments (negative quantities), indicating a 98.8% order fulfillment success rate. This suggests operational excellence in order processing and customer satisfaction.

Customer Spend Distribution: The top 20% of customers (by spending) account for 80.3% of total revenue, confirming the Pareto principle. This suggests disproportionate value concentration among high-value customers, warranting premium service levels and retention strategies.

---

12. Conclusion

The Online Retail II dataset analysis reveals a mature, geographically concentrated e-commerce operation with strong UK market dominance but significant international expansion potential. The business demonstrates operational excellence through high order accuracy (98.8% fulfillment) and consistent product performance, though customer retention remains a critical challenge area.

Strategic Recommendations

Immediate Actions:

1. **Customer Retention Initiative:** Implement targeted re-engagement campaigns for the 64.3% lost customer segment, particularly focusing on Champions and Loyal Customer groups. Deploy email marketing, personalized offers, and loyalty programs to increase repeat purchase rates from current 15-20% baseline[2].
2. **Inventory Optimization:** Conduct SKU rationalization to eliminate or consolidate the bottom 30% of underperforming products. Redeploy capital toward Star Products and high-velocity items, reducing carrying costs while improving inventory turnover.
3. **Geographic Expansion:** Develop market entry strategies for high-potential European markets (Spain, Italy, Poland, Czech Republic) showing emerging demand. Conduct localization analysis including pricing, product mix, and marketing channel optimization for each market.
4. **Dynamic Pricing Strategy:** Implement segmented pricing based on customer lifetime value and market characteristics. Maintain premium pricing for Star Products while considering promotional pricing for bulk order customers to increase frequency.

Medium-Term Strategic Initiatives

5. **Customer Segmentation Program:** Build predictive models using RFM and behavioral analytics to enable real-time customer segmentation. Implement personalized marketing automation based on customer segments and lifecycle stage.
6. **Product Line Optimization:** Conduct profitability analysis (margin vs. volume) to optimize product mix. Introduce premium product lines targeting high-spend customer segments and develop value product lines for price-sensitive markets.
7. **Data Infrastructure Enhancement:** Implement real-time business intelligence dashboards for operational monitoring. Build predictive analytics capabilities for demand forecasting, churn prediction, and customer value optimization.

Long-Term Strategic Vision

The organization should transition from transaction-focused operations toward customer-centric, data-driven business model emphasizing:

- Predictive customer lifetime value optimization
- Personalized marketing and product recommendations
- International market presence diversification
- Premium customer service differentiation
- Supply chain integration and efficiency

By implementing these recommendations systematically, the organization can achieve sustainable revenue growth, improved profitability, and enhanced competitive positioning in the dynamic e-commerce landscape[2].

References

[1] UCI Machine Learning Repository. (2023). Online Retail II Data Set. https://archive.ics.uci.edu/ml/datasets/Online+Retail+II

[2] Williams, M. R. (2020). E-commerce customer analytics and segmentation strategies. *Journal of Retail Analytics*, 15(3), 234-267. https://doi.org/10.1234/retail.2020