# Market Basket Analysis for Online Retail Dataset

**Final Comprehensive Report**

Prepared by: Skillytixs Project C Team

**Team Members -**

Data Wrangler: Santosh

Python Analyst: Sakshi

SQL Analyst: Pratheeksha

Strategy Analyst: Pranita

BI Developer + Project Lead: Kislay

## 1. Introduction

The Online Retail dataset provides a rich, real-world environment for applying data analytics, machine learning, and business intelligence techniques. This dataset includes over 540,000 transactions, covering product descriptions, quantities purchased, prices, customer identifiers, and invoice-level metadata such as time, date, and country of purchase.

This report consolidates each phase of the end-to-end analytical workflow:

- Data extraction and cleaning
- Exploratory data analysis
- Market basket analysis using Apriori algorithm
- SQL-based customer, product, and revenue insights
- Business intelligence dashboard development
- Strategic recommendations for marketing, inventory, and merchandising

The goal of the project is to support **decision-making related to customer behavior, product performance, cross-selling opportunities, and stock optimization**.

## 2. Dataset Overview

The Online Retail dataset contains:

**Key Attributes**

- **InvoiceNo** – unique transaction ID
- **StockCode** – product code
- **Description** – product name
- **Quantity** – number of units purchased
- **InvoiceDate** – date and time of transaction
- **UnitPrice** – price per unit
- **CustomerID** – unique customer identifier
- **Country** – location of customer

## Business Context

This dataset originates from a UK-based online store specializing in:

- Home décor
- Gifts
- Accessories
- Seasonal items
- Party materials

The dataset contains both B2C and B2B transactions. Bulk orders and large invoice sizes suggest a combination of wholesale and retail customers.

## Analytical Potential

Because the dataset includes product-level granularity for each invoice, it is ideal for:

- Detecting **frequently bought-together items**
- Understanding **customer purchasing journeys**
- Analyzing **sales cycles** and **seasonal trends**
- Assessing **product price variation**
- Developing **marketing bundles and promotions**

## 3. Phase 1 — Data Cleaning & Preparation (Santosh)

High-quality datasets are critical for reliable analytics. The raw dataset included inconsistencies, missing values, negative quantities, and formatting issues.

## 3.1 Handling Missing Values

- Missing **CustomerID** values were removed because they prevent customer-level analysis.
- Missing **Description** values were excluded because product identification becomes impossible without description.

## 3.2 Removing Duplicates

More than 5,000 duplicate rows were detected and removed.
Duplicate entries could distort frequency counts, revenue metrics, and association rule results.

## 3.3 Fixing Negative Quantities

Negative quantities represented **product returns**, cancellations, or adjustments.
These rows were removed when conducting *sales analysis* but retained in a separate dataset for internal operational review.

## 3.4 Standardizing Product Descriptions

Text fields contained inconsistencies such as:

- All-caps and mixed-case descriptions
- Trailing spaces
- Special symbols
- Minor spelling variations

Cleaning included:

- Lowercasing
- Stripping whitespace
- Removing non-alphanumeric characters
- Harmonizing similar products

## 3.5 Creating the TotalPrice Feature

A financial metric was engineered:

TotalPrice = Quantity × UnitPrice

This allowed for:

- Invoice revenue analysis
- Product revenue contribution
- Country-level revenue trends

## 3.6 Final Clean Dataset

After cleaning, the dataset was exported as:

**OnlineRetail_cleaned.csv**

This clean version was used for SQL queries, Apriori modeling, and Power BI dashboarding.

## 4. Phase 2 — Python Market Basket Analysis (Sakshi)

The Apriori algorithm was applied to discover patterns of co-purchased products. Market Basket Analysis helps businesses understand what customers are likely to buy together.

### 4.1 Transaction Transformation

Transactions were grouped by **InvoiceNo**, converting product lists into a binary basket format suitable for Apriori.

### 4.2 Frequent Itemsets

Using a support threshold, the model identified:

- Commonly purchased individual items
- Pairs of products frequently bought together
- Larger combinations such as gift bundles

### 4.3 Association Rules

Rules were generated using:

- **Confidence** – likelihood of purchasing item B given A
- **Lift** – how much item A increases the likelihood of purchasing item B
- **Support** – frequency of the rule across all transactions

### 4.4 Key Rules Identified

**Rule Example 1**

**White Hanging Heart T-Light Holder → Jumbo Bag Red Retrospot**

- High lift indicates *strong complementary purchase behavior*.

**Rule Example 2**

**Regency 3-Tier Cake Stand → Decorative Baking Accessories**

- Suggests a strong pattern among customers preparing for parties or events.

### 4.5 Business Interpretation

The Apriori results support:

- Cross-selling strategies
- Bundle creation
- Shelf layout optimization
- Personalized product recommendations

## 5. Phase 3 — SQL Insights (Pratheeksha)

By processing the cleaned dataset with SQL, detailed insights were extracted regarding customer behavior, product performance, and revenue trends.

Below is a structured summary of the insights reflected in SQL output.

### 5.1 Sales Performance Overview

- **Total Revenue:** £8.88 million
- **Total Quantity Sold:** 5.15 million units
- **Total Unique Customers:** 4,338

These financial KPIs form the foundation for trend analysis and strategy planning.

### 5.2 Top Products by Quantity

1. Paper craft little birdie – 80,995 units
2. Medium ceramic top storage jar – 77,916 units
3. WW2 gliders assorted designs – 54,319 units

These high-volume items represent core demand products.

### 5.3 Top Products by Revenue

1. Paper craft little birdie – £168,469
2. Regency 3-tier cake stand – £142,265
3. White hanging heart T-light holder – £100,392

These items should be prioritized for:

- Inventory replenishment
- Featured promotions

- Upsell opportunities

## 5.4 Pricing Irregularities

Substantial price variation was detected in some SKUs:

- POST postage – £8,141 variation
- M manual – £4,161 variation
- DOT postage – £1,588 variation

These issues may reflect inconsistent pricing or data-entry errors and warrant further audit.

## 5.5 Geographic Sales Insights

Revenue distribution:

- United Kingdom: **~£7.28M**
- Netherlands
- EIRE
- Germany
- France

The UK overwhelmingly dominates revenue, aligning with the retailer's home market.

## 5.6 Customer-Level Insights

- **Customer 14646** generated the highest total spending.
- **Customer 14911** purchased the largest variety of products (1,787 unique SKUs).

These high-value customers likely represent:

- Wholesale buyers
- Event planners
- Repeat purchasers

## 5.7 Time-Based Insights

- Highest sales occur between **10 AM – 1 PM**
- Strong seasonal spike in **November–December**
- Weekend purchases lean toward gifting and home décor items

## 5.8 Invoice Insights

Some invoices contained **500+ unique products**, indicating:

- Large wholesale orders
- Restocking events for small retail shops
- Corporate gifting or seasonal preparation purchases

## 6. Phase 4 — Power BI Dashboard Summary (Kislay)

The Power BI dashboard visually synthesizes findings from Python and SQL using:

- KPI cards
- Geographical maps
- Bar charts
- Line charts
- Pie charts
- Slicers
- Quarterly and monthly trends

Key dashboard insights:

- Q4 exhibits the highest revenue growth.
- Monthly trend shows strong upward trajectory toward year-end.
- UK dominates overall contribution.
- Fast-moving SKUs align with SQL results.

## 7. Phase 5 — Strategic Business Recommendations (Pranita)

Integrating insights from all analytical phases, the following recommendations were developed:

## 7.1 Inventory Optimization

- Increase stock for core products with high demand and high revenue contribution.
- Reduce inventory holding costs by scaling down low-movement, high-price items.

## 7.2 Cross-Selling Strategy

Use Apriori results to build:

- Product bundles
- Gift sets
- "Frequently bought together" suggestions
- Theme-based promotional packages (party sets, décor kits)

## 7.3 Customer Segmentation & Loyalty Strategy

High-value customers should receive:

- Personalized recommendations
- Wholesale pricing tiers
- Loyalty programs
- Early access to new product lines

## 7.4 Pricing Standardization

Audit SKUs with unexplained price variation.
Establish consistent pricing rules for:

- Discounts
- Bulk purchases
- Repeat customers

## 7.5 Seasonal Promotion Strategy

Since Q4 boosts sales significantly:

- Launch holiday campaigns earlier
- Reinforce stock levels of décor products
- Offer bundled seasonal packs

## 8. Quality Assurance (QA) Summary (Santosh)

A QA checklist was conducted across all stages:

✓ **Data Cleaning QA**

- Nulls, duplicates, and invalid quantities addressed
- Product names standardized
- Financial calculations validated

✓ **Python Apriori QA**

- Support thresholds documented
- Rule evaluation scoring validated
- Redundant rules removed

✓ **SQL QA**

- All queries tested successfully
- Revenue totals matched BI dashboard
- Customer segmentation accuracy validated

✓ **BI Dashboard QA**

- Slicers synchronized
- KPI cards match SQL outputs
- Visuals properly labeled
- Title corrected as requested

✓ **Documentation QA**

- Report structured with clear headings
- Visuals and insights linked logically
- All phases fully represented

## 9. Conclusion

This end-to-end analytical project:

- Identified key revenue-driving products
- Mapped customer buying behavior
- Revealed strong seasonal and time-based trends
- Detected cross-selling opportunities through Apriori rules
- Highlighted geographic and demographic segments
- Produced a BI dashboard for continuous monitoring
- Created actionable business strategies for marketing and operations

The project demonstrates collaborative data science execution and industry-relevant reporting standards suitable for retail analytics, supply chain optimization, and customer intelligence initiatives.