Dataset Selection Document

1. Business Domain Chosen and Why:

- **Domain:** Retail
- **Reason:** Retail offers real-world transactional data with clear ties to revenue, inventory, and customer behavior—ideal for cleaning messy data (cancellations, missing IDs) and delivering 5 actionable, revenue-focused insights.

2. Data Source(s) with Links:

- **Dataset:** Online Retail UCI Dataset
- **Source:** UCI Machine Learning Repository
- **Direct Download Link**: Download

3. Brief Description of What the Raw Data Contains

The dataset contains transactional records from a UK-based online retail company operating from December 2010 to December 2011. It includes 541,909 rows of individual item purchases and cancellations across 38 countries, primarily the United Kingdom. The data captures product, pricing, quantity, customer, and timestamp details, making it ideal for revenue, customer behavior, and inventory analysis.

Column	Description
InvoiceNo	6-digit unique invoice number; starts with 'C' for cancellations
StockCode	5–6 character product code
Description	Product name (text)
Quantity	Number of units purchased (negative = cancellation/return)
InvoiceDate	Date and time of transaction (dd/mm/yyyy hh:mm)
UnitPrice	Price per unit in GBP
CustomerID	5-digit unique customer identifier (float; many nulls)
Country	Country of customer

4. **Business questions** that are going to be solved in this assignment:

1. Which products contribute the most to total revenue, and is there a Pareto (80/20) pattern?
2. How does customer lifetime value (CLV) vary by purchase frequency and country?
3. When and why do cancellations peak (by day, hour, or product), and what revenue is lost?
4. Are there clear seasonal or monthly patterns in sales volume and average order value?
5. Which countries offer the highest profit margins, and where should marketing budget be prioritized?