Business Problem Statement (Online Retail)

A retail company wants to use its transaction data to improve revenue performance, customer engagement, and long-term loyalty. Leadership needs a clear view of what drives sales across countries, products, and time, and how many purchases are attributable to known customers versus anonymous checkout.

Overarching Business Question

> *"How can the company leverage online retail transaction data to identify revenue drivers, reduce anonymous revenue, and improve retention through segmentation and targeted actions?"*

Data Context (What we have)

Invoice-level line items with: InvoiceNo, Description, Quantity, UnitPrice, InvoiceDate, CustomerID, and Country. The dataset includes cancellations/returns (InvoiceNo starting with 'C'), non-positive quantities/prices, duplicates, and a meaningful share of orders without CustomerID (anonymous customers).

Objectives (What the business wants to know)

- **Revenue performance:** revenue, orders, and AOV by month and by country; identify seasonal peaks.
- **Concentration risk:** which customers and products contribute a disproportionate share of revenue.
- **Customer behavior:** repeat vs one-time customers; differences between UK (high volume) and non-UK (often higher AOV).
- **Retention:** segment customers using RFM and estimate churn risk (proxy: no purchase in the next 60 days).
- **Actionability:** translate insights into inventory, marketing, and retention recommendations.

Deliverables

- **Data Preparation & Feature Engineering (Python):** clean invalid sales lines and build customer RFM features.
- **Data Analysis (SQL / SQLite):** reusable business queries (revenue trends, top products/customers, AOV, repeat rate, segments).
- **Visualization (Power BI):** interactive dashboard with slicers (date, country, customer type).
- **Modeling (Churn proxy):** train a baseline + stronger model (e.g., Random Forest / XGBoost) to output churn probability.
- **Report + GitHub:** concise written summary and a well-structured repository containing notebooks, SQL, and dashboard.

Success Metrics (KPIs to report)

Revenue & Orders	Total Revenue, Total Orders (distinct invoices), AOV
Market Mix	UK revenue share, Top non-UK markets, AOV by country
Identity Capture	Anonymous revenue %, Identified customers (unique CustomerID)
Retention	Repeat customer share, Segment sizes (Champions/Loyal/At Risk/Need Attention), Churn rate (proxy)

Note: Because explicit churn labels and customer demographics are not present, churn is defined as a time-based proxy (no purchase within N days). Results are designed for decision support (who to target and when), not causal claims.