Online Retail II — Project Documentation

TEAM MEMBERS:

1- ALI ABDOU ALI
2- AMIRA MAGDY SAAD

0) Dataset Overview

Name: Online Retail II (e-commerce transactions, 20010–2011)
Grain: one row = one invoice line (product on an invoice)
Convention: InvoiceNo starting with "C" denotes a return/cancellation .

Description:

- The dataset is rich but messy.
- Contains 541,910 rows and 8 columns. It is valuable for retail analytics but requires significant preparation.
- It needs systematic cleaning (duplicates removal, invalid row filtering, text standardization, CustomerID handling).

Data Dictionary

Column	Type	Description
InvoiceNo	Text	Transaction ID. If it starts with "C", it's a return/cancellation.
StockCode	Text	Product (SKU) code; may be numeric or alphanumeric (variants).
Description	Text	Product description
Quantity	Integer	Units for the line.
InvoiceDate	Datetime	Timestamp; basis for Year/Month/Week/Hour features.
UnitPrice	Integer	Price per unit (£).
CustomerID	Integer (nullable)	Unique customer; can be null. Keep for order-level; exclude for customer-level (RFM).
Country	Text	Customer country.

Part A — SQL Server :

A1) Cleaning

- Rename some headers :
  - Invoice→InvoiceNo,
  - Price→UnitPrice,
  - CustomerID→CustomerID
- Handle Rows Where InvoiceNo Does Not Match Valid Rules (Update to Unique Random Values)
- Update Invalid Stock Codes with Random Unique Values
- Convert Null Values of Description to Unknown.
- Remove exact duplicates with a ROW_NUMBER() CTE (delete rn>1).
- Normalize text with LTRIM/RTRIM; type casts via TRY_CONVERT; convert blanks to NULL.

A2) Transformations

- **TotalPrice** = Quantity * UnitPrice
- **IsReturn** = CASE WHEN LEFT(InvoiceNo,1)='C' THEN 'Return' ELSE 'Sale' END.
- **Year/Month/YearMonth/Week/Hour** derived from InvoiceDate
- Convert the **InvoiceDate** column to a **date-only** format

A3) Example Views

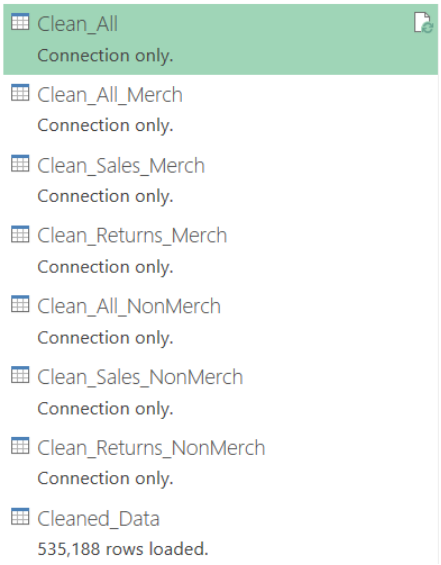
- **Helpers view** clean.vw_Sales_All: adds IsReturn, TotalPrice, YearMonth, Year, Month, Week ,Hour ,and a date cast.
- **Sales-only view** clean.vw_Sales_Valid: IsReturn='Sale' AND UnitPrice>0 AND Quantity>0 AND InvoiceDate IS NOT NULL.
- **Returns view** clean.vw_Returns: IsReturn='Return'.

A4) Example Insights (Queries)

- **Customer Retention Rate On Specific Time**(percentage of customers who have made more than one purchase over a specific period. This helps evaluate customer loyalty)
- **Customer Retention Rate At All** (percentage of customers who have made more than one purchase)
- **Days Between First and Second Purchase** (Assess the ability and time to attract New Customers)
- **Identifying Lapsed Customers** (Identifying customers who have stopped purchasing, which can guide retention strategies)
- **Best Days for Sales** (Make Offers discounts on specific day of the week, like Black Friday)
- **Top Return Rate Products** (Request feedback from customers who returned these products to understand the issues with them)
- **Sales Patterns Throughout the Day** (Increase the number of employees during peak hours)
- **Time for Repeat Purchase** (Calculating the average time it takes for a customer to make a repeat purchase, which is key for retention efforts)
- **Sales Distribution Across Categories** (What are the strong categories where I should increase stock, and which are the weak ones that I should reduce?)
- **Products Often Bought Together** (Maintain stock of these two products Available at the same time, And offer Discounts for purchasing them together.)
- **Monthly Revenue** (To Give a Bonus on salary of Employees where a revenue is High in Specific Month)
- **Best Performing Markets** (Increase shipping contracts with the most profitable countries and expand retail outlets in those regions)
- **Overall Return Rate** (Understand product satisfaction and quality issues)
- **Monthly Return Rate** (Are there any issues in Specific month with the products due to the shipping company?)
- **Top Customers (Monetary) + RFM**
  - Offer purchase installments to customers with the highest spending
  - Reward the most frequent customers by offering them special deals and fast shipping benefits
- **Volume vs Value** (Some items sell many units but generate low revenue, while others sell fewer units but generate high revenue. This insight helps evaluate sales efficiency.)
- **Order-Level KPIs (AOV & Basket Size)** (Analyzing Average Order Value (AOV) and basket size to measure sales efficiency per order)
- **Items Purchased Per Order** (Tracking the number of items purchased per order to evaluate the purchasing behavior)
- **Sales Efficiency Per Order** (Measuring sales efficiency for each order to assess performance)

Part B — Excel (Power Query + Pivots)

The cleaning has been implemented across many different queries to avoid deleting rows and to accurately represent the insights with true factual measurements



B1) Cleaning in Power Query

- Rename some headers :
  - Invoice→InvoiceNo,
  - Price→UnitPrice,
  - CustomerID→CustomerID
- Remove Whole Rows Duplicates
- Trim & Clean text , Upper/Proper Case.
- Filter sales view: UnitPrice > 0 and Quantity > 0 , IsReturn=Sale.
- Convert CustomerID Nulls → Unknown.
- CustomerID Unknown: keep for order-level; filter Unknown for customer-level.

B2) Transformations

- **Custom Column**: IsReturn = if Text.Start([Invoice],1)="C" then "Return" else "Sale".
- Derive **Year**, **Month**, **YearMonth**, **Week of Year** ,**Weekday**, **Hour** from InvoiceDate.
- **TotalPrice** = [Quantity] * [Price]/[UnitPrice].
- **NonMerch_Merch :** To Trace the Non-Product Lines
  - POST → **POSTAGE** → **1,253** rows → **£66,248.64**
  - DOT → **DOTCOM POSTAGE** → **709** rows → **£206,245.48**
  - M → **Manual** → **571** rows → **–£68,674.19**
  - C2 → **CARRIAGE** → **143** rows → **£6,986.00**
  - D → **Discount** → **77** rows → **–£5,696.22**
  - S → **SAMPLES** → **63** rows → **–£3,049.39**
  - BANK CHARGES → **Bank Charges** → **37** rows → **–£7,175.64**
  - AMAZONFEE → **AMAZON FEE** → **34** rows → **–£221,520.50**
  - 23574 → **PACKING CHARGE** → **16** rows → **£90.00**
  - gift_0001_10/20/30/40/50 → **Dotcomgiftshop Gift Voucher £xx** → **31** rows total → **~£686**
  - CRUK → **Commission → 16** rows → **£7933.43**

B2) Example Insights (PivotTables & Charts)

- **Sales | Returns Comparaison For Most Losing Products** (The Problem is related to Sales or Returns)
- **Sales | Returns Comparaison For Top Revenue Products** (what is the side effect of Returns)
- **Ratio of Sales For NonMerch&MerchProducts**(To know the real Sales related to our business evaluation)
- **Ratio of Returned Invoices** (Evaluate customer satisfaction with the product.)
- **Average items per order over Months** (Need to increase the production of Stock in a specific month)
- **Monthly Revenue** (To Give a Bouns on salary of Employees where a revenue is High in Specfic Month)
- **Sales vs Revenue vs Returns over months** (Is my profit margin good?)
- **Returns over months** (Are there any issues in Specific month with the products due to the shipping company?)
- **Best WeekDays Sales** (Make Offers discounts on specific day of the week, like Black Friday)
- **Sales Over Day Hours** (Increase the number of employees during peak hours)
- **Top Products Sales** (working with types of products similar to the best-sellers)
- **Top Products Revenue** (Increase the stock of the most Saled products)
- **Top Customers Spend** (Offer purchase installments to customers with the highest spending)
- **Top Customers had Invoices** (Reward the most frequent customers by offering them special deals and fast shipping benefits)
- **Top Countries Revenue** (Increase shipping contracts with the most profitable countries and expand retail outlets in those regions)
- **Most Returned Products** (Request feedback from customers who returned these products to understand the issues with them)
- **Most Returning Countries** (Review the factories in those countries, the quality of the retail outlets, and the shipping companies)

1) Before vs After (What Changed)

Handle **duplicate lines**; sales rows with **UnitPrice ≤ 0** or **Quantity ≤ 0**.
Kept **Returned Invoices separate**: return lines (InvoiceNo starts with 'C').
Added: IsReturn, TotalPrice, Year/Month/YearMonth/Week/Weekday/Hour.
Standardized: trimmed Description; removed stray symbols (e.g., leading '*'); unified casing.
CustomerID nulls: kept for order-level; excluded for customer-level (RFM/top customers).

2) Deliverables Checklist

SQL: table + views + analysis queries (monthly revenue, country revenue, returns, top products, AOV, basket size).
Excel: Power Query pipeline; PivotTables + charts; Dashboard with slicers (Year, Country, IsReturn).

3) Notes for the Write-up

Invoices repeat because each row is a line item; for order-level KPIs group by InvoiceNo.
Sales analysis excludes UnitPrice ≤ 0 and Quantity ≤ 0; returns analyzed separately.
CustomerID nulls are retained for order-level analysis but excluded from customer-level (RFM).
RFM scores use quantiles for Recency, Frequency, Monetary to segment customers.