Online Retail Revenue & Customer Behavior Analysis

1. Project Overview

This project analyzes an Online Retail transactional dataset to understand revenue drivers, customer purchasing patterns, and retention opportunities. The workflow covers data auditing and cleaning (Python), insight extraction (SQL in SQLite), customer segmentation (RFM), and a business dashboard (Power BI).

2. Dataset Summary

- Dataset type: invoice-level transaction lines (products, quantities, unit prices, dates, countries,customers).
- Primary tables used for analysis:
• **transactions_cleaned**: cleaned transactions for revenue/product/country trends
• **customer_rfm**: customer-level RFM features for segmentation and churn proxy modeling

3. Key Data Quality Findings (Python Audit)

From quick checks on the raw dataset:
• Duplicate rows detected: **5,268**
• Unique invoices: **25,900**
• Unique identified customers (CustomerID present): **4,372**

Quality flags affecting revenue reliability:
• Cancellation rows (InvoiceNo starts with 'C'): **9,288**
• Negative/zero quantity rows: **10,624**
• Non-positive unit price rows: **2,517**

Industry insight: a meaningful portion of transactions are not linked to CustomerID (anonymous checkout). For customer-level analytics (RFM/churn), anonymous rows are excluded; for transaction-level revenue/product analysis they can be retained.

4. Data Cleaning & Feature Engineering (Python)

Two clean datasets were created because business questions differ by level of analysis:

Transaction-level analysis (trends, products, countries): keep all valid sales lines to maximize coverage.
Customer-level modeling (RFM + churn proxy): use only identified customers and valid sales to avoid leakage and incorrect features.

Customer modeling filter rules (applied before building RFM):

✓ CustomerID is not null

✓ InvoiceNo does not start with 'C' (exclude cancellations/returns)

✓ Quantity > 0

✓ UnitPrice > 0

RFM features engineered:

- **Recency**: days since the customer's last purchase (smaller = more recent)
- **Frequency**: number of unique invoices (orders) per customer • **Monetary**: total revenue contribution per customer (sum of Sales)

Outputs saved for reuse across tools:

- **data/processed/transactions_cleaned.csv** for SQL/BI
- **data/processed/customer_rfm.csv** for segmentation + modeling

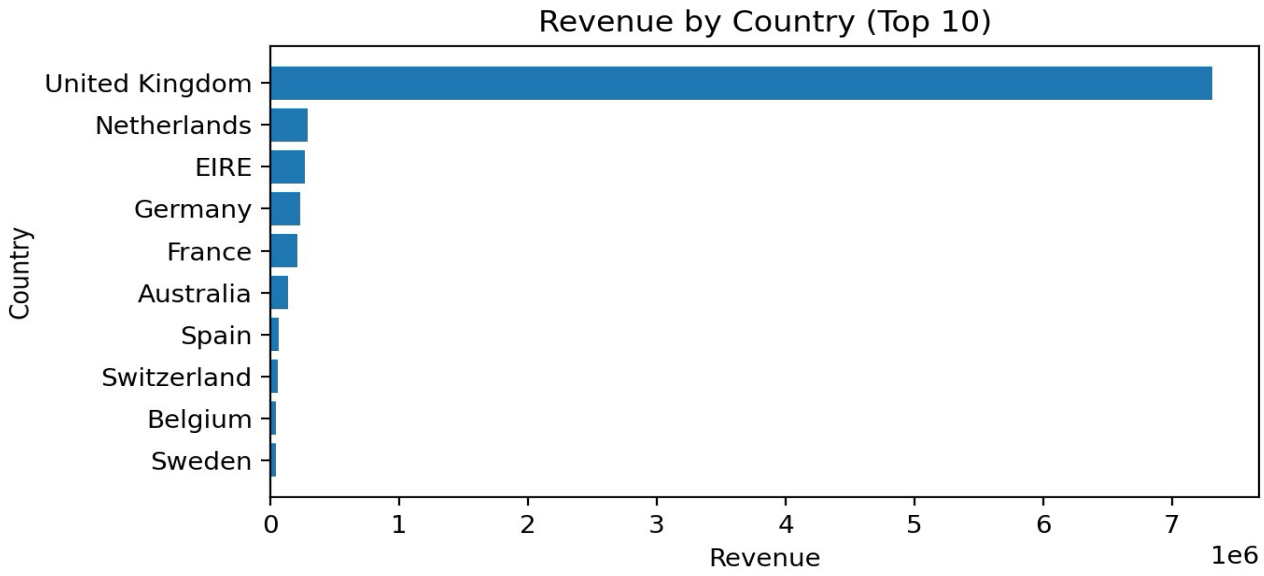
5. Revenue Insights (SQL in SQLite)

Headline KPIs (cleaned transactions):

• Total revenue: **8,911,407.90**
• Total orders (distinct invoices): **18,532**
• Overall AOV (revenue / orders): **480.87**
• UK revenue share: **82.0%**

Q1. Revenue by Country (Top 10)

Country	Revenue	Orders	Customers
United Kingdom	7,308,391.55	16,646	3,920
Netherlands	285,446.34	94	9
EIRE	265,545.90	260	3
Germany	228,867.14	457	94
France	209,024.05	389	87
Australia	138,521.31	57	9
Spain	61,577.11	90	30
Switzerland	56,443.95	51	21
Belgium	41,196.34	98	25
Sweden	38,378.33	36	8

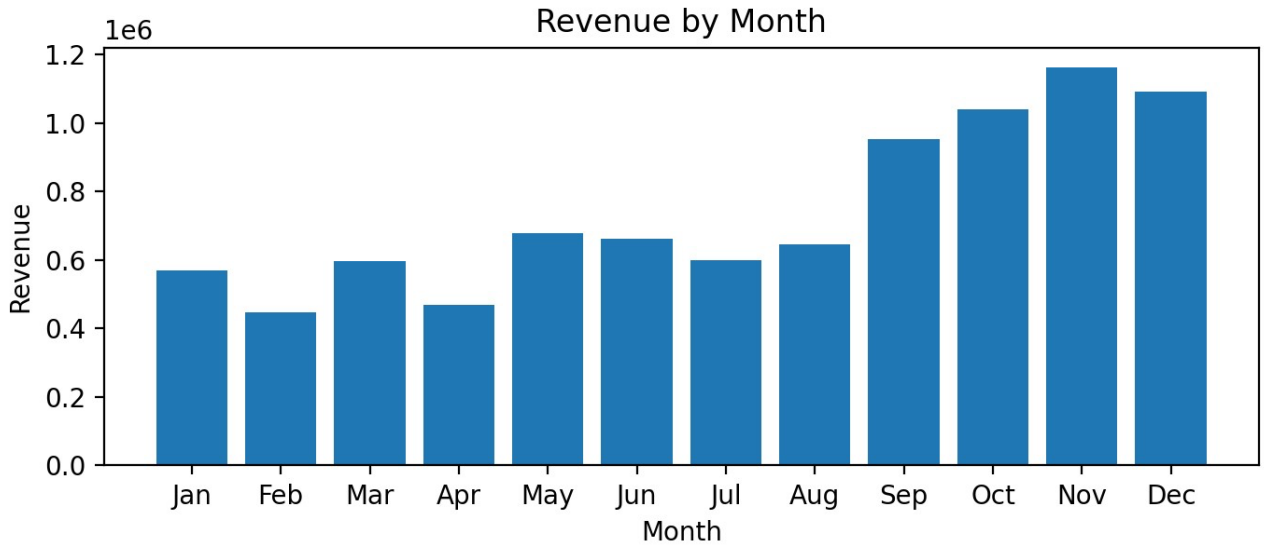


Q2. Monthly Revenue Trend

Revenue shows clear seasonality: it builds toward the end of the year, with the highest sales typically in Oct–Dec (peak in Nov).

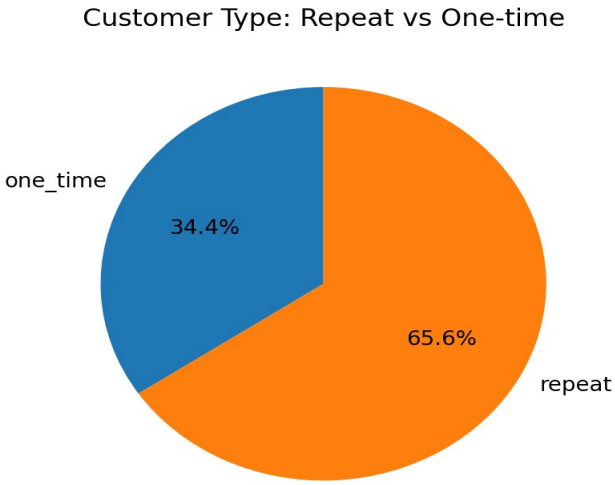
Month	Revenue	Orders
Jan	569,445.04	987
Feb	447,137.35	997
Mar	595,500.76	1,321
Apr	469,200.36	1,149
May	678,594.56	1,555
Jun	661,213.69	1,393

Jul	600,091.01	1,331
Aug	645,343.90	1,280
Sep	952,838.38	1,755
Oct	1,039,318.79	1,929
Nov	1,161,817.38	2,657
Dec	1,090,906.68	2,178



Q4. Repeat Purchase Proxy

Repeat customers (≥2 orders) are 65.6% of the customer base, while 34.4% buy only once. This is a clear retention opportunity.


Customer Type: Repeat vs One-time
one_time
34.4%
65.6%
repeat

Q3. Top Customers by Revenue

Revenue is concentrated among a small set of high-value customers. Losing a few top customers can noticeably impact total sales.

CustomerID	Revenue	Orders	Avg Line Value
14646	280,206.02	73	134.97
18102	259,657.30	60	602.45
17450	194,550.79	46	577.30
16446	168,472.50	2	56,157.50
14911	143,825.06	201	25.34
12415	124,914.53	21	174.95
14156	117,379.63	55	83.84
17511	91,062.38	31	94.56
16029	81,024.84	63	334.81
12346	77,183.60	1	77,183.60

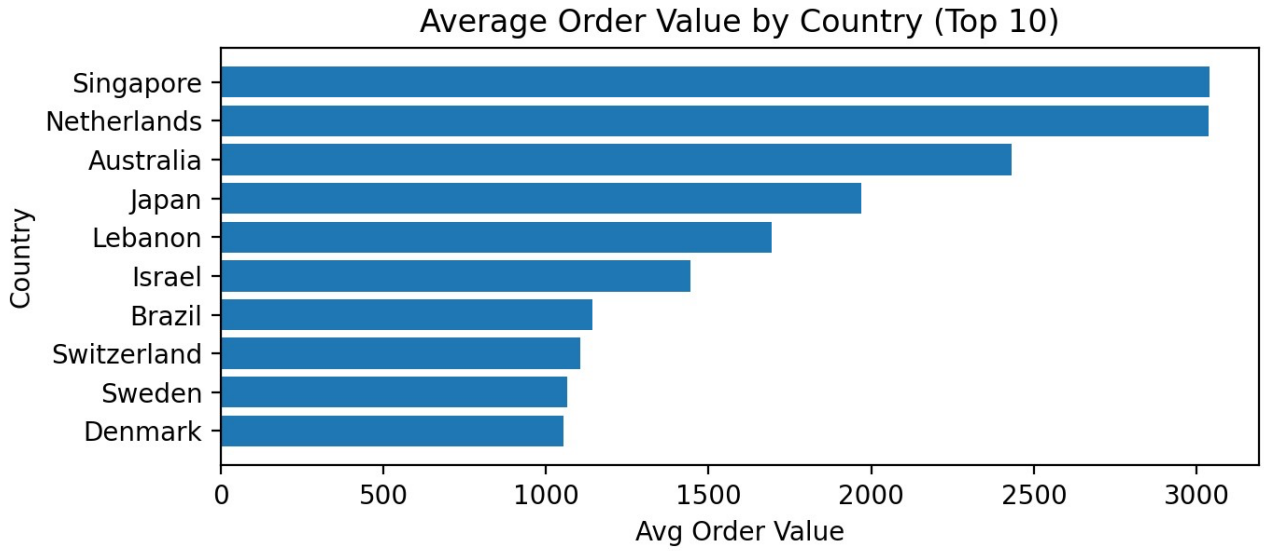
Q5. Top Products by Revenue (Top 10)

StockCode	Description	Revenue	Units Sold
23843	PAPER CRAFT , LITTLE BIRDIE	168,469.60	80,995
22423	REGENCY CAKESTAND 3 TIER	142,592.95	12,402
85123A	WHITE HANGING HEART T-LIGHT HOLDER	100,448.15	36,725
85099B	JUMBO BAG RED RETROSPOT	85,220.78	46,181
23166	MEDIUM CERAMIC TOP STORAGE JAR	81,416.73	77,916
POST	POSTAGE	77,803.96	3,120
47566	PARTY BUNTING	68,844.33	15,291
84879	ASSORTED COLOUR BIRD ORNAMENT	56,580.34	35,362
M	Manual	53,779.93	7,173
23084	RABBIT NIGHT LIGHT	51,346.20	27,202

Q6. Average Order Value (AOV) by Country

UK generates the highest total revenue because it has many more orders and customers, but its average order value is lower than many international markets.

Country	Avg Order Value
Singapore	3,039.90
Netherlands	3,036.66
Australia	2,430.20
Japan	1,969.28
Lebanon	1,693.88
Israel	1,444.34
Brazil	1,143.60
Switzerland	1,106.74
Sweden	1,066.06
Denmark	1,053.07

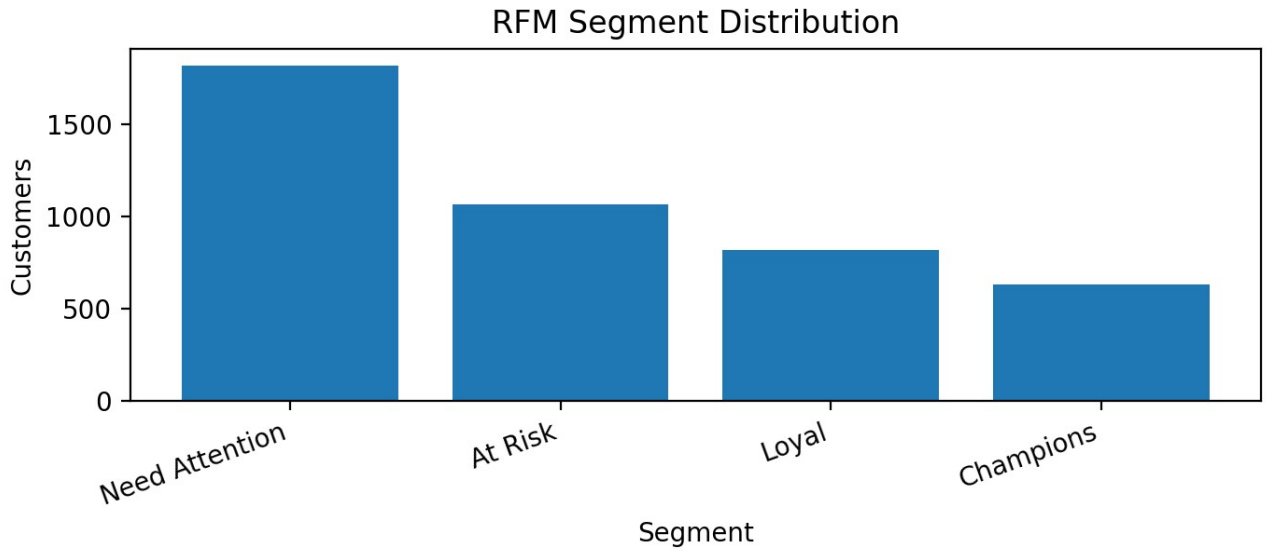


Interpretation: International orders are higher-ticket (high AOV) but low volume; UK is high volume but lower AOV. Business can grow by (1) improving UK basket size via bundles/upsells, and (2) scaling high-AOV markets that already show strong order value.

6. Customer Segmentation (RFM)

RFM segmentation was built for **4,338** identified customers after removing cancellations and invalid sales lines.

Segment	Customers	Share	Avg Recency (da	s)Avg Frequency	Avg Monetary
Need Attention	1,820	42.0%	81.9	2.88	1,126.25
At Risk	1,065	24.6%	217.9	1.10	487.71
Loyal	820	18.9%	20.2	5.18	2,440.42
Champions	633	14.6%	5.9	12.41	6,857.94



Actions by segment:
• **Champions**: VIP perks, early access, protect from churn (highest frequency & monetary).
• **Loyal**: cross-sell and upsell, subscription/loyalty program nudges.
• **Need Attention**: reminders and personalized offers to increase frequency.
• **At Risk**: win-back campaigns (they haven't purchased for ~218 days on average).

7. Churn Proxy Modeling (Random Forest & XGBoost)

Because the dataset does not include an explicit churn flag, churn was defined as: **no purchase in the next 60 days** after a chosen cutoff date. RFM features were computed using only the history window to avoid future leakage.

Model	Accuracy	ROC-AUC	PR-AUC
RandomForest	0.61	0.668	0.663
XGBoost	0.66	0.721	0.711

Prediction delivered by the model: a churn probability for each customer. Operations teams can target high-risk customers with win-back offers, while protecting Champions with retention benefits.

8. Power BI Dashboard (Final Delivery)

The dashboard brings together KPIs, trends, and customer segments for business stakeholders. Recommended slicers: date range, country, and customer type (identified vs anonymous).

Business Recommendations (Industry-ready)

1) **Improve identity capture** (email/account/loyalty) to reduce anonymous revenue and enable retention/personalization.

2) **Plan for seasonality**: allocate inventory and marketing budget ahead of Sep–Dec peak. 3) **Protect hero products**: prevent stockouts for top revenue items and bundle complementary products to lift UK AOV.

4) **Retention focus**: win-back At Risk customers and grow Need Attention customers into Loyal via targeted offers.

5) **International expansion**: prioritize high-AOV markets (e.g., Netherlands, Singapore, Australia) with tailored shipping/marketing.