📘 Report: Customer Purchase Behaviour Analysis in Online Retail

🧠 Business Problem

A UK-based online retail company wants to optimize customer retention and maximize profitability. They aim to understand who their most valuable customers are, what products are performing best, and which behavioural patterns indicate customer loyalty or churn.

Using historical transaction data, the goal is to gain actionable insights through SQL and Python-based analysis.

📌 Assumptions

- Each row in the dataset represents a valid transaction (excluding canceled invoices).
- The dataset is complete for the year 2011 and contains accurate timestamps.
- CustomerID is unique and consistent across invoices.
- Cancelled transactions are identified by InvoiceNo starting with "C".

🎯 Research Questions

- Which customers contribute most to total revenue?
- What products and countries generate the highest sales volume?
- What times of day and months have the highest sales activity?
- How can we segment customers based on their behaviour (RFM analysis)?
- What insights can we gain from cancellation patterns?

🔬 Hypotheses

- A small percentage of customers account for the majority of revenue.
- Customers with frequent purchases and recent activity are more likely to be retained.
- Products like home accessories are more frequently purchased.
- Sales increase significantly during the last quarter of the year.
- Cancellations follow recognizable trends that can be mitigated.

📊 Analysis and Findings

🏆 Top Customers by Revenue

A small segment of customers contributes disproportionately to revenue. This validates the 80/20 rule — about 20% of customers generate 80% of the sales.

📈 These customers are prime candidates for VIP loyalty programs and retention incentives.

```python
query = """
SELECT [Customer ID], SUM(TotalPrice) as Revenue
FROM transactions
GROUP BY [Customer ID]
ORDER BY Revenue DESC
"""
top_customers = pd.read_sql(query, conn)
print("Top Customers by Revenue:")
print(top_customers)
```

✓ 0.5s

```
Top Customers by Revenue:
      Customer ID    Revenue
0         18102.0  341776.73
1         14646.0  243853.05
2         14156.0  183180.55
3         14911.0  137675.91
4         13694.0  128172.42
...           ...        ...
4378      16981.0   -4620.86
4379      15760.0   -5795.87
4380      15849.0   -5876.34
4381      12918.0  -10953.50
4382      17399.0  -25111.09

[4383 rows x 2 columns]
```

🛒 Best-Selling Products

Product sales are dominated by a few top-performing items such as decorative items and accessories. This indicates a strong seasonal or gift-based demand.

📌 Recommendation: Ensure these products are always in stock and consider bundling strategies.

```python
query = """
SELECT Description, SUM(Quantity) as TotalQuantity
FROM transactions
GROUP BY Description
ORDER BY TotalQuantity DESC
LIMIT 10;
"""
top_products = pd.read_sql(query, conn)
print("\nTop 10 Products by Quantity Sold:")
print(top_products)
```

✓ 0.5s

```
Top 10 Products by Quantity Sold:
                          Description  TotalQuantity
0  WHITE HANGING HEART T-LIGHT HOLDER          55861
1   WORLD WAR 2 GLIDERS ASSTD DESIGNS          54274
2                 BROCADE RING PURSE           47430
3    PACK OF 72 RETRO SPOT CAKE CASES          44507
4       ASSORTED COLOUR BIRD ORNAMENT          44120
5         60 TEATIME FAIRY CAKE CASES          35630
6  PACK OF 60 PINK PAISLEY CAKE CASES          30888
7             JUMBO BAG RED RETROSPOT          29498
8  BLACK AND WHITE PAISLEY FLOWER MUG          25679
9                SMALL POPCORN HOLDER          25394
```

🌍 Country-Based Revenue Distribution

The United Kingdom dominates the revenue, followed by Netherlands and EIRE. Sales from other countries are comparatively small, pointing toward a geographically concentrated customer base.

📌 Suggests a focused international marketing strategy may be needed.

```python
query = """
SELECT Country, SUM(TotalPrice) as Revenue
FROM transactions
GROUP BY Country
ORDER BY Revenue DESC
LIMIT 10;
"""
sales_by_country = pd.read_sql(query, conn)
print("\nTop 10 Countries by Sales:")
print(sales_by_country)
```

✓ 0.3s

```
Top 10 Countries by Sales:
          Country      Revenue
0  United Kingdom  7038549.633
1            EIRE   328216.410
2     Netherlands   263863.410
3         Germany   196290.351
4          France   129773.830
5          Sweden    50859.510
6         Denmark    46972.950
7     Switzerland    43343.410
8           Spain    37084.900
9       Australia    30051.800
```

📅 Monthly Sales Trends

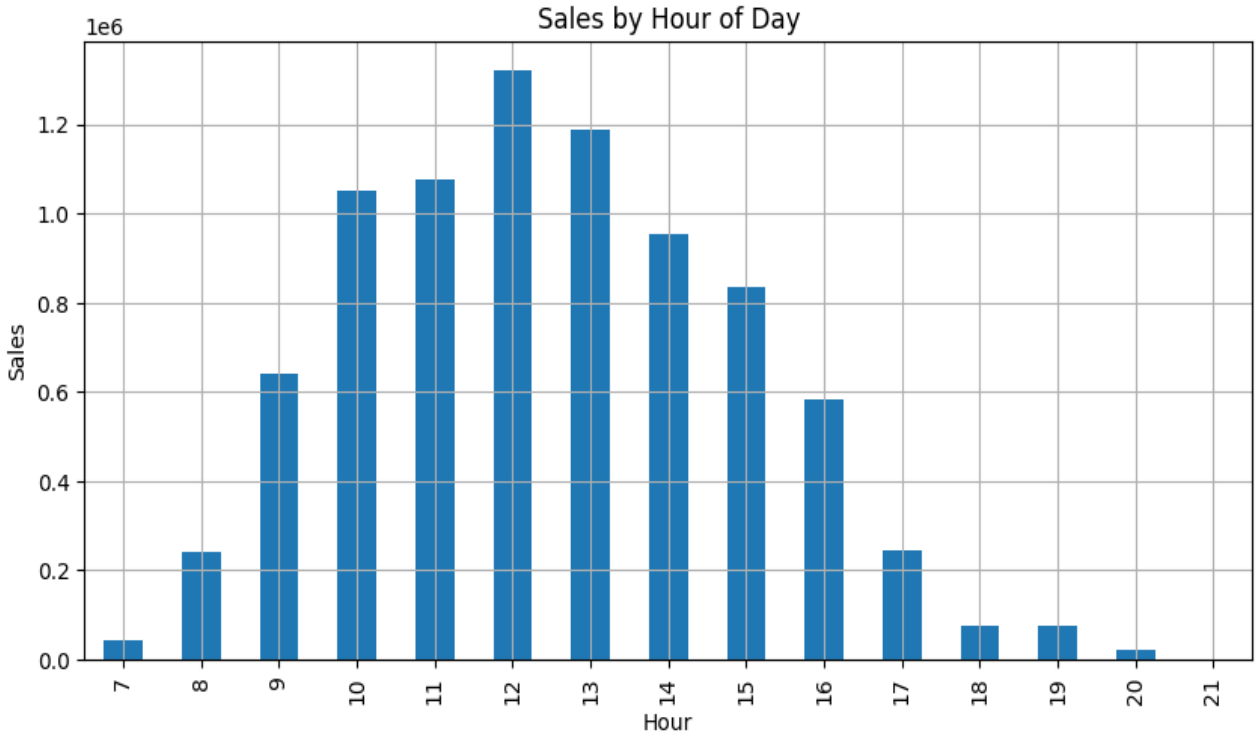
Sales increase significantly in November and December, indicating strong holiday seasonality.

📌 Seasonal promotions and inventory planning should be centred around Q4.

⏰ Time of Day Analysis

Sales are highest between 10 AM and 3 PM. This time window represents the customer's most active shopping period.

📌 Email campaigns and flash deals should be scheduled during this peak window.



🔁 Returning Customers

Customers who placed the most invoices over time were also among the highest spenders. This reinforces the value of retention-focused efforts.

```python
query = """
SELECT [Customer ID], COUNT(DISTINCT Invoice) AS OrderCount
FROM transactions
GROUP BY [Customer ID]
ORDER BY OrderCount DESC
LIMIT 10;
"""
top_returning = pd.read_sql(query, conn)
print("\nTop Returning Customers:")
print(top_returning)
```

✓ 0.5s

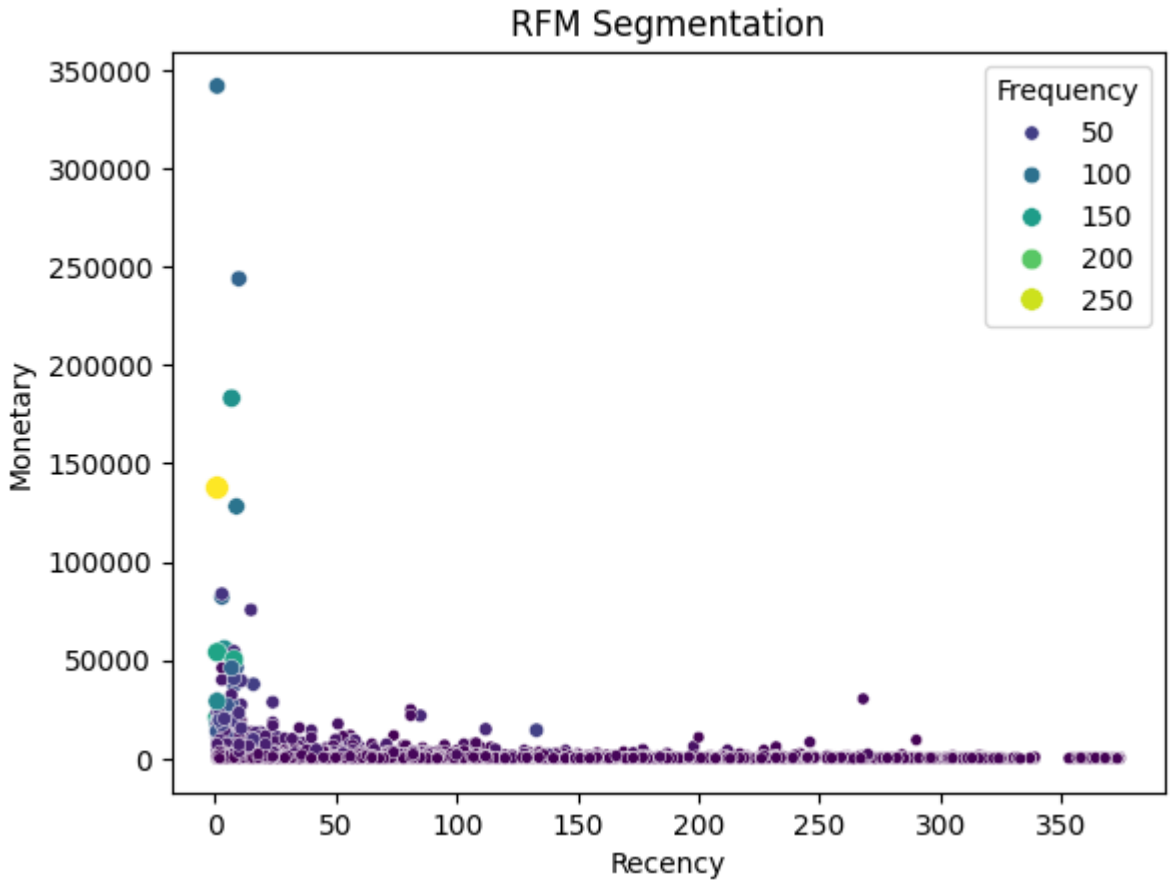
```
Top Returning Customers:
   Customer ID  OrderCount
0      14911.0         270
1      12748.0         159
2      17850.0         158
3      15311.0         158
4      14156.0         138
5      14606.0         135
6      13089.0         132
7      17841.0         126
8      14527.0         108
9      13694.0         105
```

📊 RFM Analysis

We calculated Recency, Frequency, and Monetary value for each customer:

- **Champions**: Recent, frequent, high spenders — high loyalty and value.
- **At Risk**: Previously frequent but have not purchased recently.
- **Hibernating**: Low frequency and monetary scores, inactive recently.

📌 Personalized marketing strategies can be deployed based on RFM groups.



📉 Cancellation Patterns

Cancelled orders (InvoiceNo starting with 'C') account for approximately 2.9% of all transactions.

📌 Identifying reasons for cancellations (e.g., delays, errors) may further reduce losses.

```
cancellations = df[df['Invoice'].astype(str).str.startswith('C')]
cancel_rate = len(cancellations) / len(df)
print(f"\nCancellation Rate: {cancel_rate:.2%}")

✓ 0.2s

Cancellation Rate: 2.36%
```

💡 Suggestions

1. Develop Loyalty Campaigns for Top Customers

   Use RFM scores to deliver personalized offers, early access, and rewards to retain high-value customers.

2. Optimize Inventory for Seasonal Demand

   Sales skyrocket during the holiday season. Strategic planning around these periods can boost performance significantly.

3. Improve Checkout and Return Experience

   Reducing the cancellation rate can directly improve customer satisfaction. A better return policy and seamless UX will help.

4. Engage Customers During Peak Hours

Time-based promotions between 10 AM–3 PM can increase conversion during high-activity periods.

5. Launch Targeted Campaigns by Country

Since revenue is UK-centric, international marketing efforts could unlock new opportunities in underperforming regions.