IDENTIFYING CUSTOMER SEGMENTS AND OPTIMIZING PRODUCT RECOMMENDATION USING RFM SEGMENTATION, COHORT ANALYSIS AND MARKET BASEKT ANALYSIS.

A CASE STUDY

Paritosh S. Ghimire

March 2025

INTRODUCTION

In today's competitive e-commerce landscape, understanding customer behaviour and product performance is no longer optional—it's essential. Businesses that fail to analyse their customers' purchasing patterns risk ineffective marketing strategies, low retention rate, and missed revenue opportunities. With thousands of transactions occurring daily, the challenge lies in extracting meaningful insights from raw data to enhance customer engagement, optimize product offerings, and drive sales growth.

The client is a medium-sized online retail business operating primarily in the United Kingdom while serving neighbouring countries as well. With an estimated annual revenue of **£7-10 million**, the company specialises in selling a wide range of consumer goods, including home decor, accessories, and personalised gifts. The business has built a strong customer base across multiple regions, leveraging e-commerce platforms to drive sales.

However, the company has been facing significant challenges which has impacted profitability and inventory management. The retailer seeks to gain a deeper understanding of consumer behaviour, improve customer segmentation, and identify commonly purchased product combinations to enhance cross-selling and bundling strategies.

Problem Identification

The online retailer faced several key challenges that were affecting both revenue and operational efficiency. These challenges stemmed from a lack of structured data insights, inefficient product bundling, high customer churn rate, and difficulty in customer retention.

Key Challenges

Limited Customer Insights:

The company lacked a well-defined approach to analysing customer purchasing patterns. Without a clear understanding of who their most valuable customers were or what products performed best across different segments, they were unable to optimise marketing and retention strategies effectively.

Inefficient Product Bundling & Cross-Selling:

The retailer was uncertain about which products were frequently purchased together. This lack of clarity resulted in missed opportunities to introduce product bundles, strategic discounts, and better cross-selling strategies. Without market basket analysis, the business was unable to take advantage of complementary product relationships to drive higher basket values.

Customer Retention & Churn Issues:

A large proportion of customers made only one or two purchases before becoming inactive. The company lacked insights into why customers were not returning and how different customer cohorts behaved over time. This issue was leading to an increasing cost per acquisition (CPA) as marketing efforts were focused on acquiring new customers rather than nurturing existing ones.

Data Quality & Structure Issues:

The raw dataset provided for analysis contained inconsistencies, missing values, and duplicate entries. This made it difficult to perform meaningful analysis without extensive data cleaning and preprocessing. Without a structured approach to data management, the business struggled to extract accurate insights that could be used to drive decision-making.

Business Challenges

- Missed opportunities for higher revenue through ineffective bundling strategies.
- High customer churn rates leading to stagnating customer lifetime value (CLV).
- Inefficient inventory management due to unclear demand patterns.

To tackle these issues, I proposed a comprehensive data analytics framework using Recency, Frequency, and Monetary (RFM) segmentation and Cohort Analysis to understand consumer behaviour. Additionally, Market Basket Analysis to identify products frequently brought together. These methods allowed us to uncover customer behavioural patterns, improve retention strategies, and optimise product bundling to increase sales.

Methodology

To ensure a structured and data-driven approach, I followed a multi-stage methodology, beginning with understanding data description & data migration, data preprocessing, data normalisation, exploratory data analysis (EDA), and application of advanced analytical techniques.

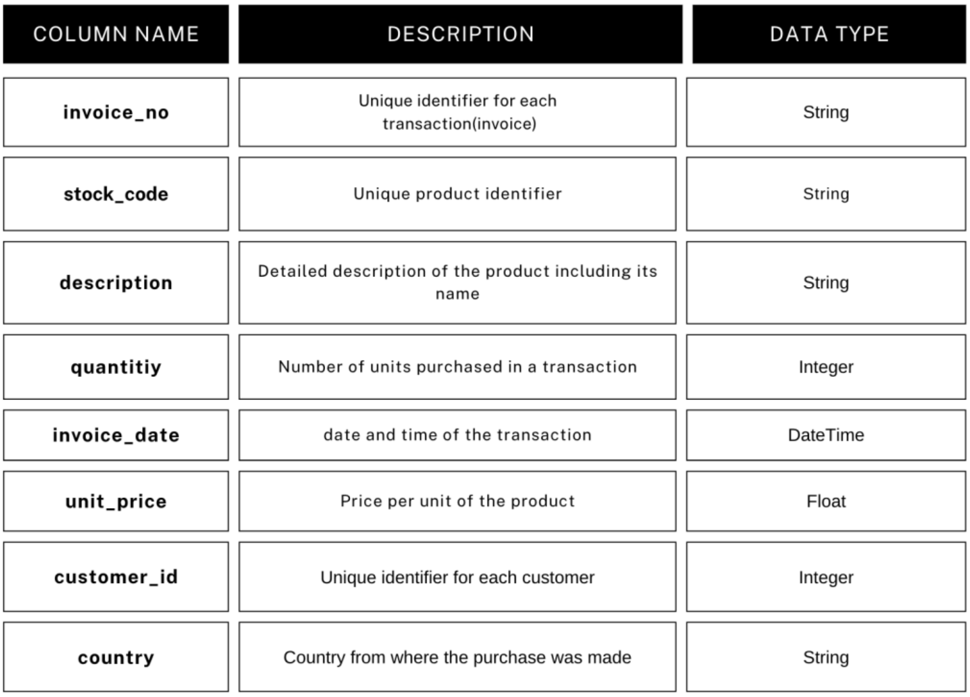
Data Description & Migration

The dataset used in this study is the Online Retail Dataset from the UCI Machine Learning Repository which includes transactions recorded between December 2010 and December 2011

Number of Records: 541,909

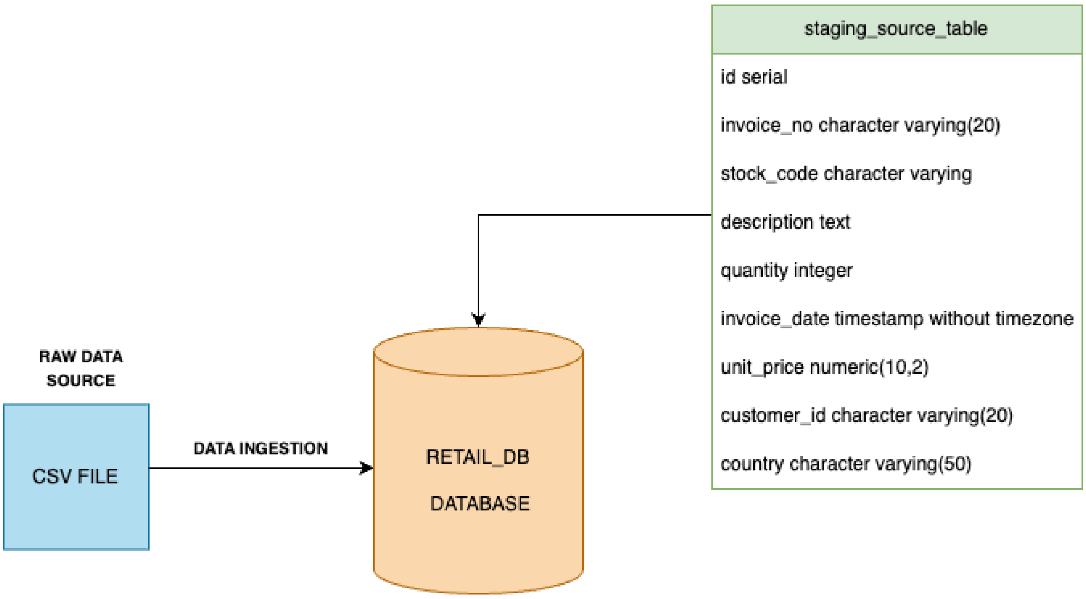
Data Granularity: Includes details at the invoice-item level (each row represents a product purchased in a transaction)

Key Columns and Definitions:



COLUMN NAME	DESCRIPTION	DATA TYPE
invoice_no	Unique identifier for each transaction(invoice)	String
stock_code	Unique product identifier	String
description	Detailed description of the product including its name	String
quantitiy	Number of units purchased in a transaction	Integer
invoice_date	date and time of the transaction	DateTime
unit_price	Price per unit of the product	Float
customer_id	Unique identifier for each customer	Integer
country	Country from where the purchase was made	String

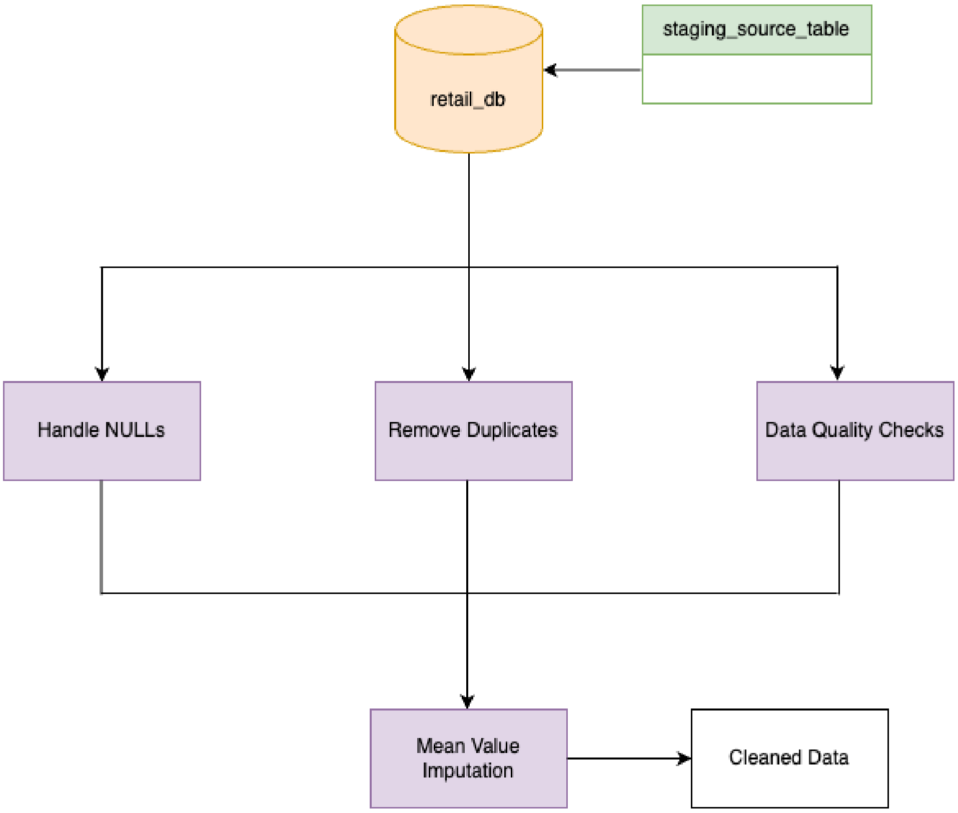
The raw data is imported into a staging table in our SQL database for initial processing. A total of 541,909 transaction rows were inserted.



staging_source_table

id serial
invoice_no character varying(20)
stock_code character varying
description text
quantity integer
invoice_date timestamp without timezone
unit_price numeric(10,2)
customer_id character varying(20)
country character varying(50)

Data Preprocessing



Upon loading the dataset to the staging table, we conducted a rigorous data cleaning and transformation process to ensure data integrity and consistency. The key steps included:

Checking for Null Values

An initial assessment of missing data revealed:

- The Description column contained 1,454 null values.
- The Customer ID column contained 135,080 null values (which was a significant concern).

Action:

Imputed nulls with placeholder '00000' to maintain consistency. The large portion of this missing value for customer id column is due to manual entries.

Data Quality Checks

Handling Duplicate Rows:

- Identified and removed 4,900 duplicate rows to prevent skewed analysis.

Handling Quantity Set to Negative:

- Found 10,587 transactions with negative quantities.
- Negative quantities were primarily linked to cancelled orders, identified by the invoice ID prefix 'C'.

Action:

Added a new column, **is_cancelled**, to easily filter out cancelled orders while preserving data integrity.

Outcome:

No negative quantity values remained except for cancelled transactions.

Handling Unit Prices Set to Zero:

- Identified 2,516 rows with unit price set to 0.
- Cancelled orders were expected to have unit price set to zero, but 1,180 rows were identified as non-cancelled orders with zero unit price.
- Further investigation revealed:
    - Most zero-priced unit transactions belonged to customer IDs set to '00000'.
    - Many of these transactions were adjustments, bank charges, or manual corrections with stock codes such as 'M' (manual), 'B' (bad debt adjustments), and 'BANK CHARGES' which were not associated with a particular product but were part of the total transaction.

Action:

- Dropped 18 rows with inconsistent stock codes and zero unit prices, likely due to data entry errors.
- Removed transactions containing irrelevant descriptions such as 'amazon', 'found', '?', 'check', and 'dotcom' that did not correspond to valid product or invoice.
- Checked for product prices that could be imputed with mean values.
- 1003 rows having stock codes with missing prices were matched with their mean unit prices and imputed accordingly.
- Remaining 20 rows could not be imputed, as they represented single-unit products with zero prices. Hence, they were dropped.

Outcome:

No more zero unit prices remained except for cancelled orders.

Handling Missing Descriptions:

- Null values in Description were not critical for our analysis and were imputed with "No Description Available" where necessary.

  Outcome:
  No more Null values in Description Column.

Handling Inconsistent Stock Codes

The stock_code column contained inconsistencies in format. After consultation with stakeholders, it was confirmed that stock codes were five-digit numbers followed by an optional alphabet suffix.

- 2770 rows with inconsistent stock codes with the format.
- Although they were inconsistent in format, they were associated with unique and valid invoices.

  Action:

  No modifications were made to stock codes, as some were linked to manual updates, bank charges, and postage fees, which were part of valid transaction types.

Final Data Validation:

- Ensured no remaining missing or inconsistent values in key columns affecting analysis.
- Verified the dataset was now clean, structured, and ready for advanced analytics.

The link to SQL data validation code can be found here

Data Normalisation

To ensure efficient data storage, eliminate redundancy, and improve query performance, the dataset was normalized into a relational database schema consisting of four well-structured tables.

The raw transactional data from the staging table was decomposed into separate entities, ensuring a one-to-many relationship between key tables while preserving data integrity.

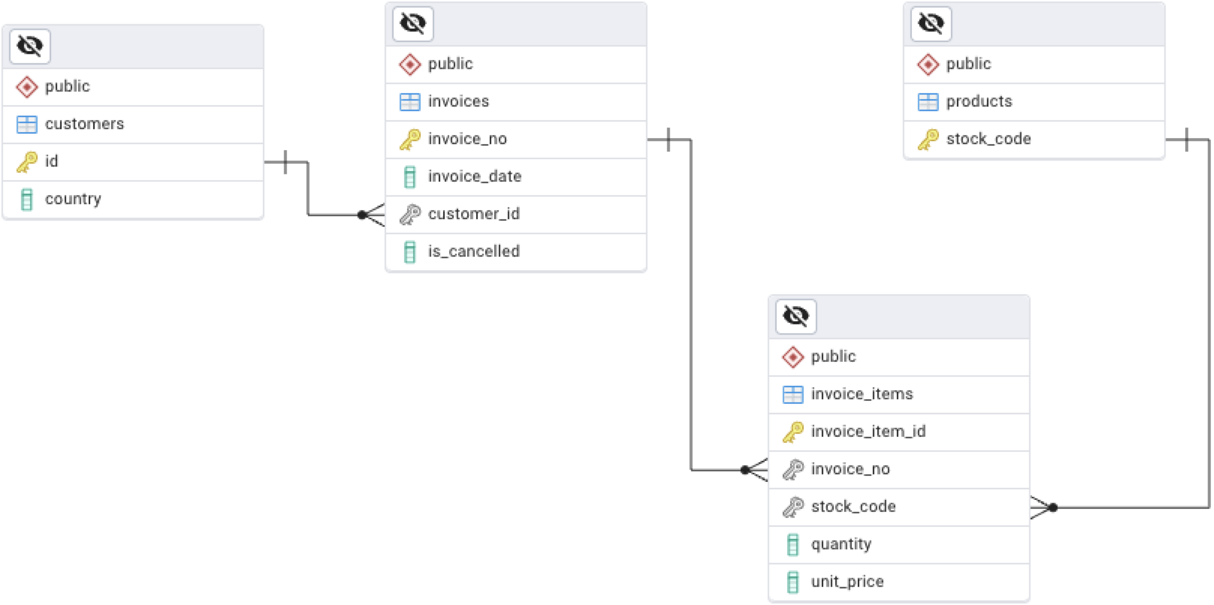
Customers Table: This table stores unique customer information using customer_id as the primary key. The country field is included for customer segmentation and regional analysis.

Invoices Table: Each transaction is recorded in this table, with invoice_no as the unique identifier. It captures details such as invoice_date and the customer_id associated with the purchase, enabling trend analysis and cohort-based customer tracking.

Products Table: This table contains product details, where stock_code serves as the primary key. It includes attributes such as unit_price for accurate pricing calculations. The description field from the raw dataset was removed to avoid redundancy, as product names were not essential for numerical analysis.

Invoice Items Table: Acting as a bridge table between invoices and products, this table records each purchased item in an invoice. It includes invoice_no, stock_code, quantity, and total_price, ensuring a clear representation of individual product purchases per transaction.

The link to SQL data normalisation code can be found here



Exploratory Data Analysis

Before performing advanced customer segmentation and market analysis, an Exploratory Data Analysis (EDA) was conducted to understand key patterns, detect anomalies, and extract meaningful insights from the dataset. This step helps in identifying business trends and customer purchasing behaviours.

The link to EDA & Trend Analysis SQL code can be found here

Summary Statistics:

A total of **4373** unique customers were identified.

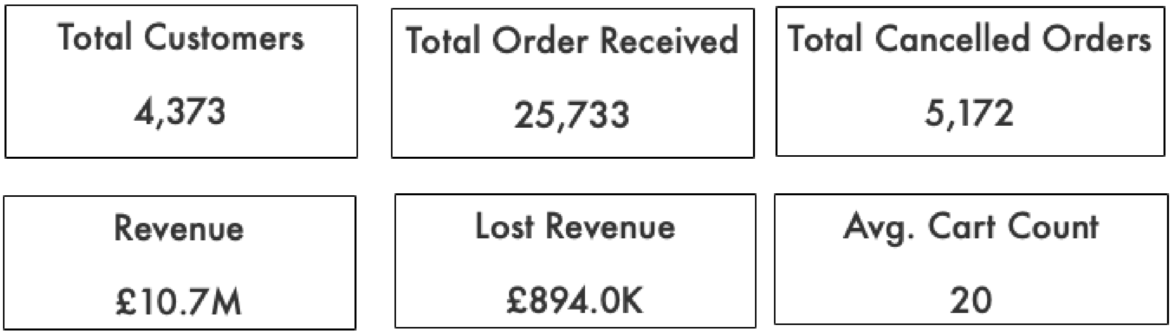
The retailer sold a total of **4052** different products in 36 different countries.

A total of **25,733** orders were received between 01/12/2010 and 09/12/2011.

Out of which, **5172** were cancelled i.e. 20% of the total order received were cancelled.

The total revenue generated is **10,643,627.27(£10.64 Million)**.

Total revenue lost due to cancellation is **£893,939.72.**



Total Customers
4,373

Total Order Received
25,733

Total Cancelled Orders
5,172

Revenue
£10.7M

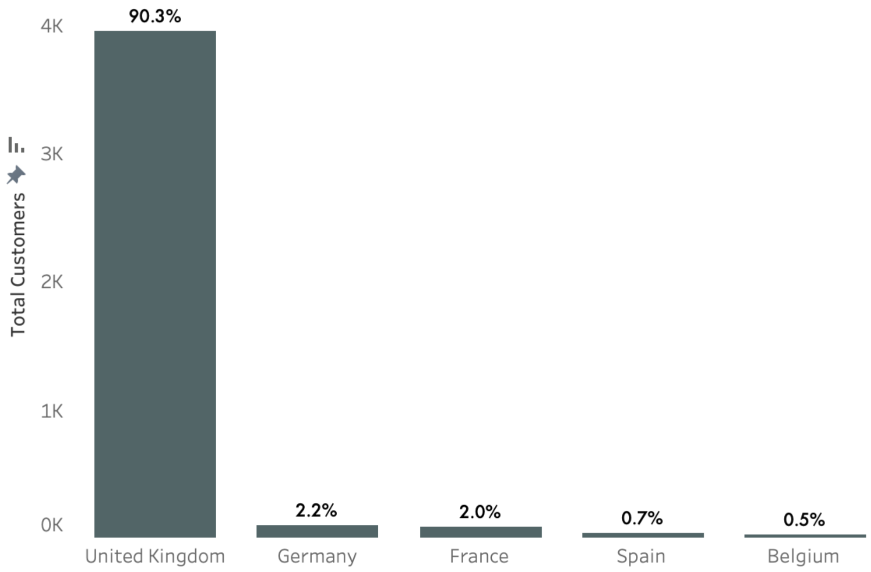
Lost Revenue
£894.0K

Avg. Cart Count
20

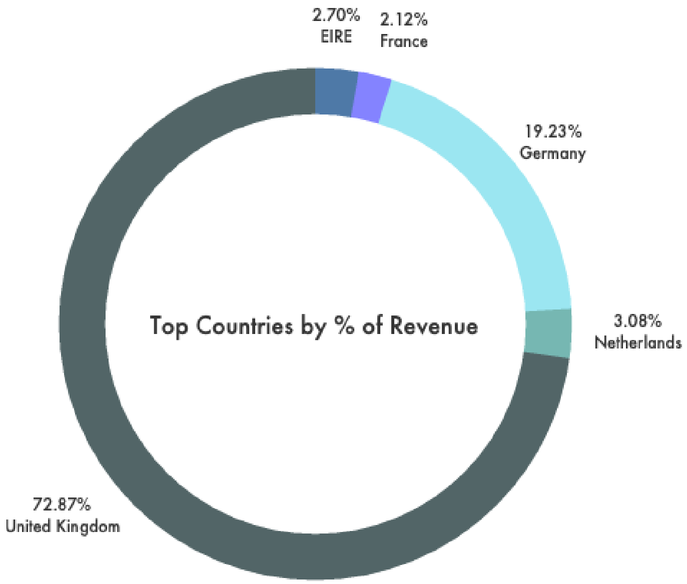
A maximum of 1114 products were ordered in a single transaction. On average, 20 products were ordered in a single transaction. 90% of total sales occurred in the United Kingdom.

Other 35 countries contributing to the remaining 10% of the orders.

The top five countries were determined based on the number of transactions, helping to highlight key markets.



The top five countries were determined based on total contribution to revenue

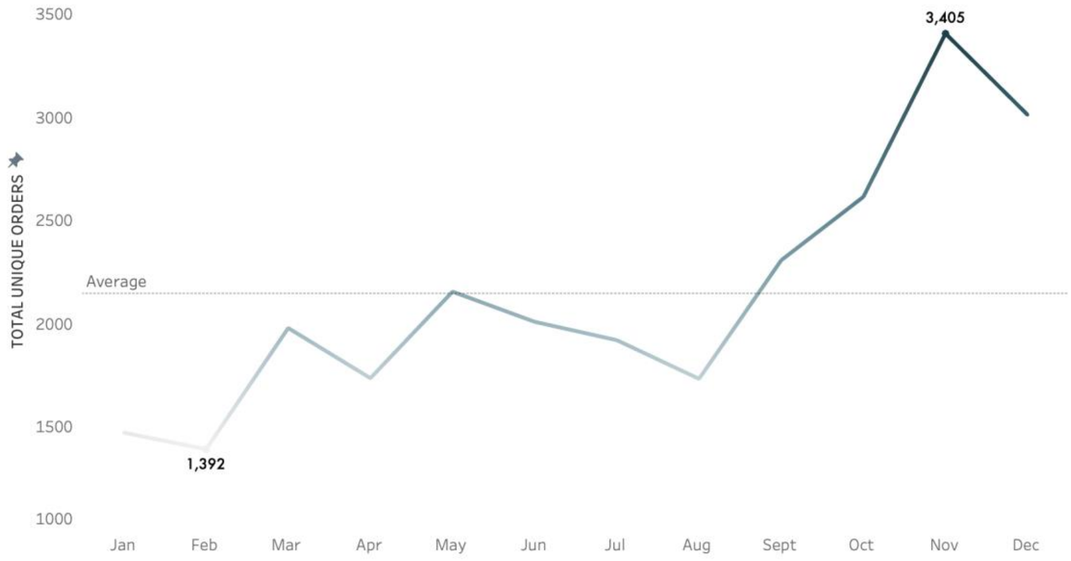


Trend Analysis

The link to EDA & Trend Analysis SQL code can be found here

Orders: Month by Month Trend Analysis

The trend of total unique orders per month reveals key insights into customer purchasing behaviour throughout the year.



Seasonal Growth in Orders

The year starts with relatively lower order volumes, with January and February showing the lowest number of unique orders (~1,392 in February). From March onwards, there is a noticeable increase in order volume, suggesting marketing campaigns driving customer engagement.

Fluctuations in Mid-Year Orders

Between April and August, order volumes stabilize with minor fluctuations, indicating a consistent demand but no significant spikes.

The slight dip in July and August suggests slowdown in customer activity.

Strong Growth in the Last Quarter

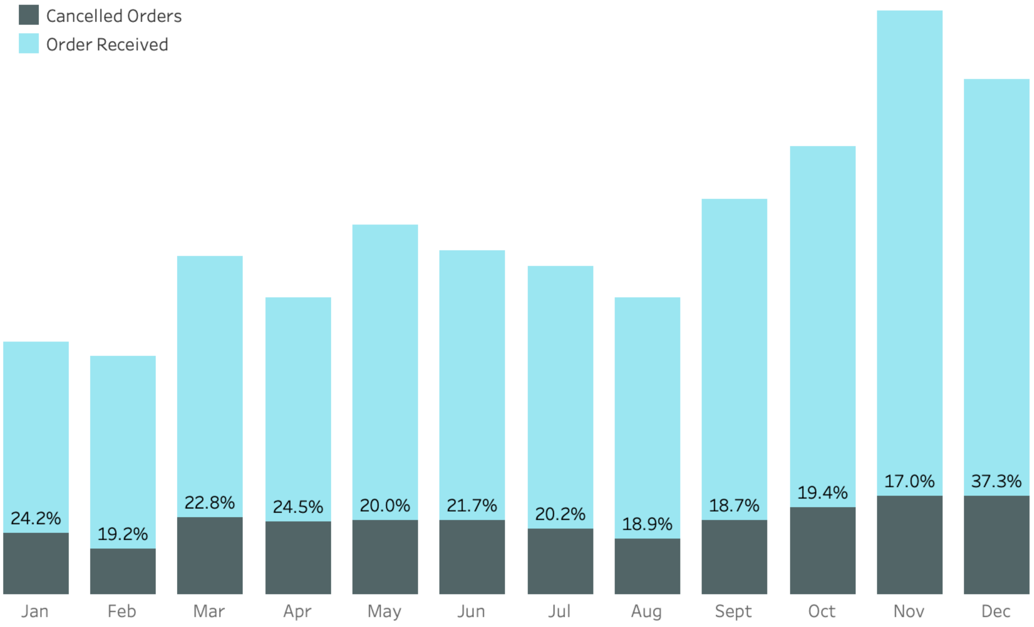
A steady increase from September to November shows a strong upward trend in customer purchases, likely driven by holiday shopping, promotional events, or seasonal demand.

November peaks at 3,405 unique orders, the highest volume of the year, which may coincide with Black Friday or pre-holiday sales events.

A slight drop follows in December is due to the cut-off on 9th of December, but order volume remains above the annual average, indicating sustained holiday-driven activity.

Cancellations: Month by Month Trend Analysis

The stacked bar chart illustrates the monthly order cancellations as a percentage of total orders, providing insights into fluctuations and trends in cancellation rates throughout the year.



General Cancellation Trend

The average cancellation rate hovers around 20% for most months.

December stands out with a significantly higher cancellation rate of 37.3%, which is nearly double the rate observed in most other months even though the data is cut-off on 9th of the month.

The lowest cancellation rates occur in August (18.9%) and November (17.0%), suggesting improved customer satisfaction or fewer order disruptions during these periods.

Seasonal & Business Impact

Q1 (Jan–Mar): Higher cancellation rates in January (24.2%) and March (22.8%), possibly due to post-holiday returns or order modifications after peak shopping in December.

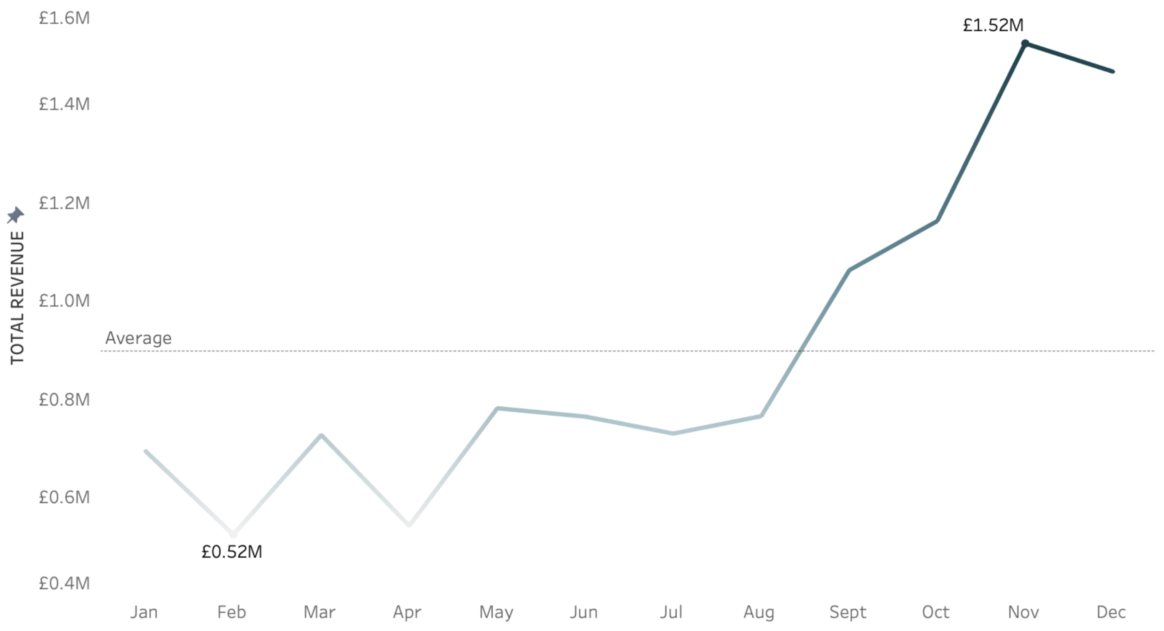
Q2 (Apr–Jun): April records the highest cancellations in this quarter (24.5%), while May and June show slight improvements (~20-21%).

Q3 (Jul–Sep): Cancellation rates are relatively stable and low, with August (18.9%) and September (18.7%) recording some of the best performance in terms of minimal cancellations.

Q4 (Oct–Dec): October & November maintain low cancellation rates (~17-19%), likely due to strong pre-holiday purchasing confidence. December shows a massive spike (37.3%), likely driven by holiday season returns, failed deliveries, or order changes post-Black Friday sales.

Revenue: Month by Month Trend Analysis

The total revenue trend closely follows the order received trend, confirming a strong correlation between order volume and revenue generation.



Low revenue in Q1 (Jan–Feb) aligns with lower order volumes, with February recording the lowest revenue (~£0.52M), consistent with the dip in unique orders.

Mid-year fluctuations (Apr–Aug) in revenue match the stable order trends, showing consistent but moderate sales activity.

A steep increase from September to November, peaking at £1.52M in November, mirrors the surge in orders, likely driven by pre-holiday and Black Friday shopping.

December revenue declines slightly despite high order volume, which may be due to higher cancellations, discounts, or refunds, aligning with the spike in December order cancellations (37.3%).

Seasonal Revenue Shifts

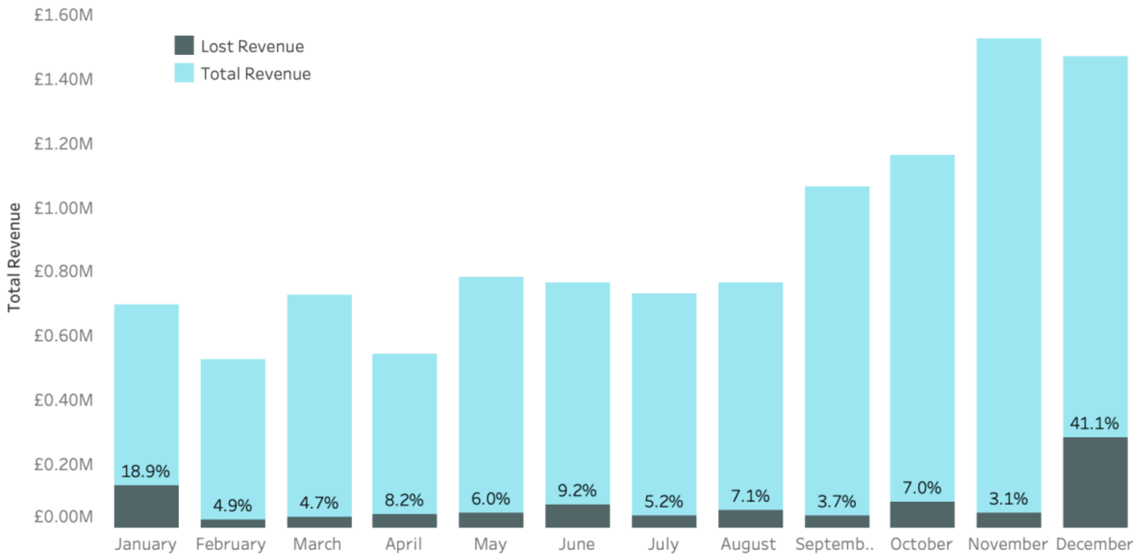
Q1 (Jan–Mar): Revenue fluctuates with low order volume and moderate cancellations (19-24%), leading to lower earnings (~£0.52M in Feb).

Q2–Q3 (Apr–Aug): Revenue stabilizes (~£0.8M) despite consistent order trends, indicating stable demand with minor seasonal dips.

Q4 (Sep–Nov): Revenue surges from September to November, peaking at £1.52M in November—the highest of the year, driven by holiday sales & Black Friday demand. November's low cancellation rate (17%) ensures strong revenue retention.

Lost Revenue: Month by Month Trend Analysis

The stacked bar chart illustrates total revenue vs. lost revenue (due to cancellations and returns), highlighting the impact of lost sales on overall earnings.



Seasonal Revenue Loss Patterns

Q1 (Jan–Mar): January (18.9%) has high lost revenue, likely due to post-holiday returns. February and March improve significantly (~5%).

Q2–Q3 (Apr–Sept): Lost revenue remains within a manageable 4-9% range, suggesting a stable period with minimal cancellations.

Q4 (Oct–Dec): November has the lowest lost revenue (3.1%), despite peak order volumes—indicating strong pre-holiday purchases with fewer returns. December's lost revenue skyrockets to 41.1%, offsetting the holiday sales boom.

Likely causes post-Christmas returns, failed deliveries, and refund requests.

Business Impact

- December's high lost revenue (41.1%) needs urgent attention.
- Despite the cut-off on 9th of December, the 41.5% of lost revenue is very concerning.

Actionable Insights

- It is essential to strengthen return policies to minimize post-holiday losses.
- Implement better customer engagement post-purchase (e.g., email follow-ups, incentives for store credit instead of refunds).
- November's low lost revenue (3.1%) suggests successful execution.
- Replicate November's sales strategies for other months. Q2–Q3 is the most stable period (4-9% loss), making it an ideal time for price testing, new product launches, and promotional experimentation.

RFM Segmentation

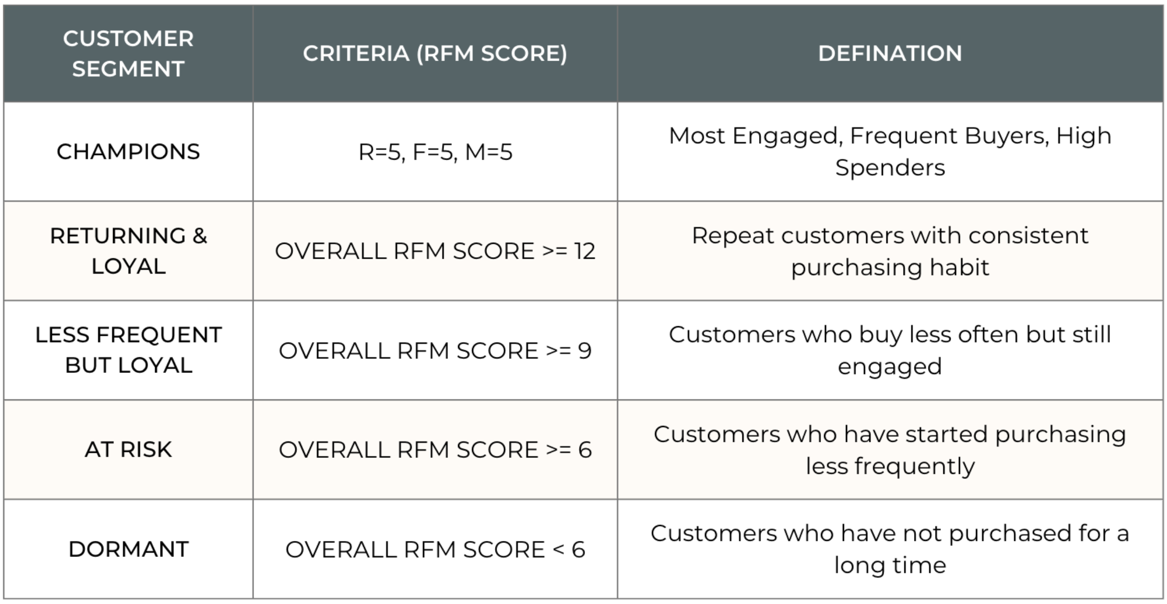
RFM segmentation is a customer value analysis technique that classifies customers based on three key user behaviour metrics:

- **Recency (R):** How recently a customer made a purchase.

- **Frequency (F):** How often they make purchases.
- **Monetary Value (M):** How much they spend overall.

By assigning a score from 1 to 5 for each metric, customers are segmented into meaningful groups for targeted marketing and retention strategies.

The link to SQL RFM Segmentation code can be found here



CUSTOMER SEGMENT	CRITERIA (RFM SCORE)	DEFINATION
CHAMPIONS	R=5, F=5, M=5	Most Engaged, Frequent Buyers, High Spenders
RETURNING & LOYAL	OVERALL RFM SCORE >= 12	Repeat customers with consistent purchasing habit
LESS FREQUENT BUT LOYAL	OVERALL RFM SCORE >= 9	Customers who buy less often but still engaged
AT RISK	OVERALL RFM SCORE >= 6	Customers who have started purchasing less frequently
DORMANT	OVERALL RFM SCORE < 6	Customers who have not purchased for a long time

The RFM segmentation analysis resulted in five key customer groups, each representing distinct purchasing behaviours.

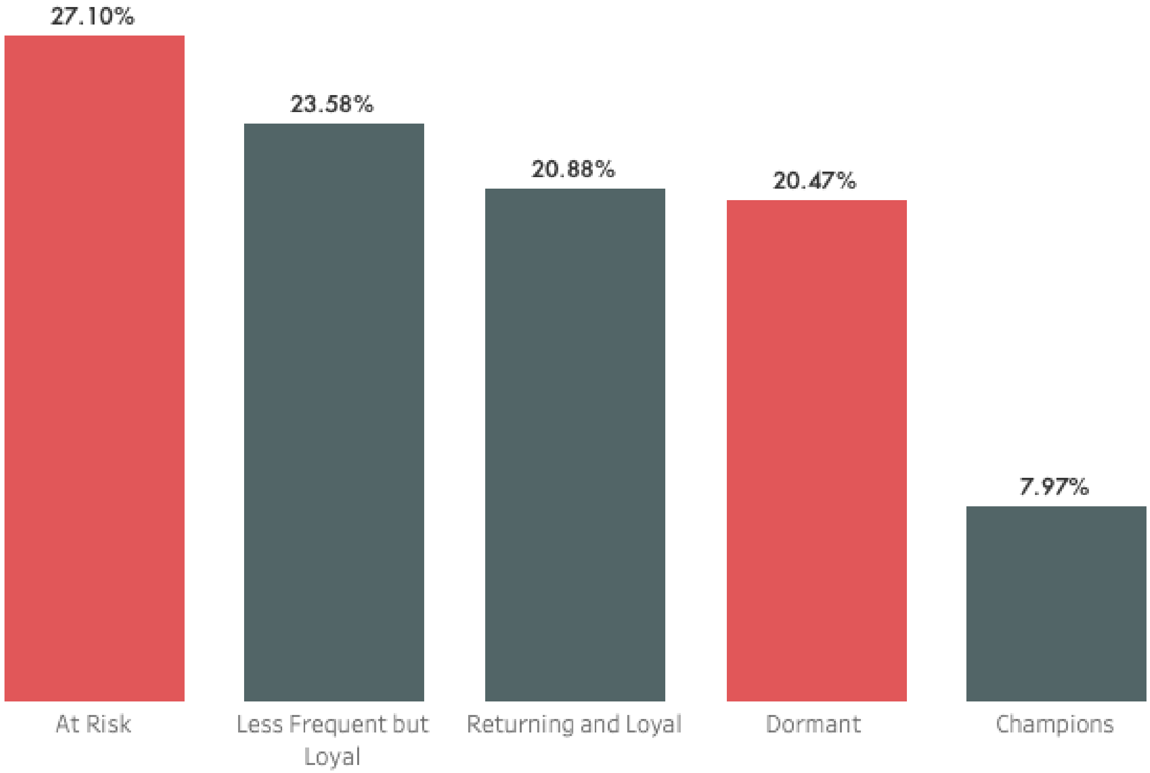
Champions (7.97%) – This segment includes 346 high-value customers who purchase frequently and spend the most. These are the most engaged buyers, making them prime candidates for exclusive offers, loyalty programs, and premium experiences to ensure continued satisfaction.

Returning and Loyal (20.88%) – Comprising 906 customers, this group consists of consistent, repeat buyers who may not spend as much as Champions but remain reliable. Personalized recommendations, membership discounts, and early access to sales can help strengthen their loyalty.

Less Frequent but Loyal (23.58%) – This segment consists of 1,023 customers who purchase less often but remain engaged. While they don’t buy frequently, they still return. Time-sensitive promotions, re-engagement emails, and special discounts can encourage them to shop more often.

Dormant (20.47%) – With 888 customers, this group includes previously active buyers who have not made a purchase in a long time. To prevent them from fully churning, businesses should implement win-back campaigns, abandoned cart emails, or special reactivation offers.

At Risk (27.10%) – This is the largest segment, with 1,176 customers, who are showing signs of disengagement. They have stopped purchasing as frequently and are on the verge of churn. Urgent interventions, personalized outreach, and targeted discounts are necessary to bring them back before they are lost completely.



Business Strategies Based on RFM Segmentation

Maximize Loyalty:

Champions & Returning Customers: Offer VIP perks, personalized discounts, and premium experiences.

Increase Purchase Frequency:

Less Frequent Buyers: Use time-sensitive offers and reminder emails to drive repeat purchases.

Re-engage & Retain:

Dormant & At-Risk Customers: Win-back strategies, abandoned cart emails, and customer service follow-ups.

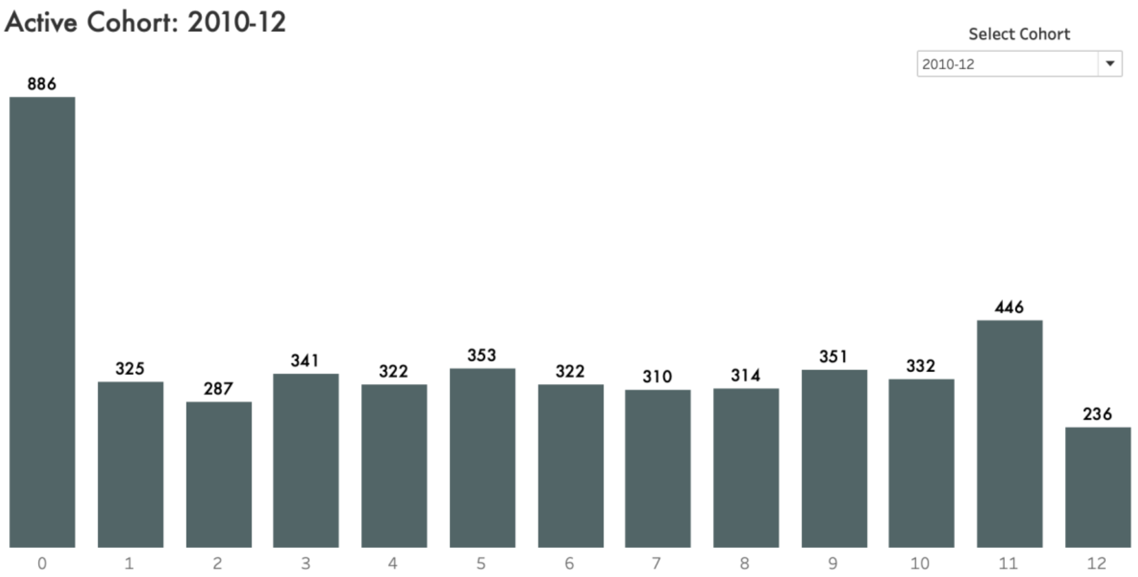
Cohort Analysis and Customer Churn

Cohort analysis tracks how groups of customers (cohorts) behave over time after their first purchase.

Methodology for Cohort Analysis:

1. Identifying the first purchase date for each customer.
2. Grouping customers into monthly cohorts based on when they made their first purchase.
3. Calculating how many months have passed since their first purchase Counting the number of active customers per cohort each month to measure retention.

The link to SQL code for Cohort Analysis can be found here



The cohort analysis shows a significant drop-off in customer retention within the first 2-3 months. This early churn pattern suggests that many customers do not return after their initial purchase, which can be caused by several factors such as:

- Lack of engagement post-purchase
- Dissatisfaction with product quality or service
- Better deals from competitors
- Lack of incentive to return

Key Insights from the Cohort Data

Customer retention declines over time, which is a common trend in most businesses.

The December 2010 cohort started with 886 customers, but by the 12th month, only 236 remained active, meaning a 73.4% drop in engagement.

Newer cohorts show similar patterns, with a significant drop in the second month.

The January 2011 cohort started with 417 customers, but by month 1, only 92 remained active (a 77.9% drop).

Most cohorts lose over 50% of their customers by the second or third month, indicating early churn is a major challenge.

Some spikes in activity (e.g., December 2010 – month 11, 446 active customers) suggest that seasonal factors or promotions temporarily bring back old customers.

Later cohorts (e.g., August–December 2011) have smaller initial numbers, which could mean slowing customer acquisition towards the end of the dataset.

Business Strategies

Improve Customer Retention Early On

First 2 months are critical—most customers drop off quickly.

Implement personalized onboarding emails, targeted discounts, and post-purchase engagement to keep new customers interested.

Offer incentives for repeat purchases within the first 30-60 days.

Leverage Seasonal Re-Engagement

The spike in December 2010 (month 11) suggests seasonal promotions work well to bring back old customers.

Further In-depth analysis of Causes of Churn in Month 2-3

Why do so many customers drop off early?

Conduct customer feedback surveys to understand drop-off reasons and fix potential issues.

Use retargeting ads & email reminders to recover lost customers in months 2-3.

To diagnose and reduce early churn, businesses should implement structured data collection beyond transaction records.

Strengthen Long-Term Retention

Customers who stay beyond 6 months tend to remain engaged longer.

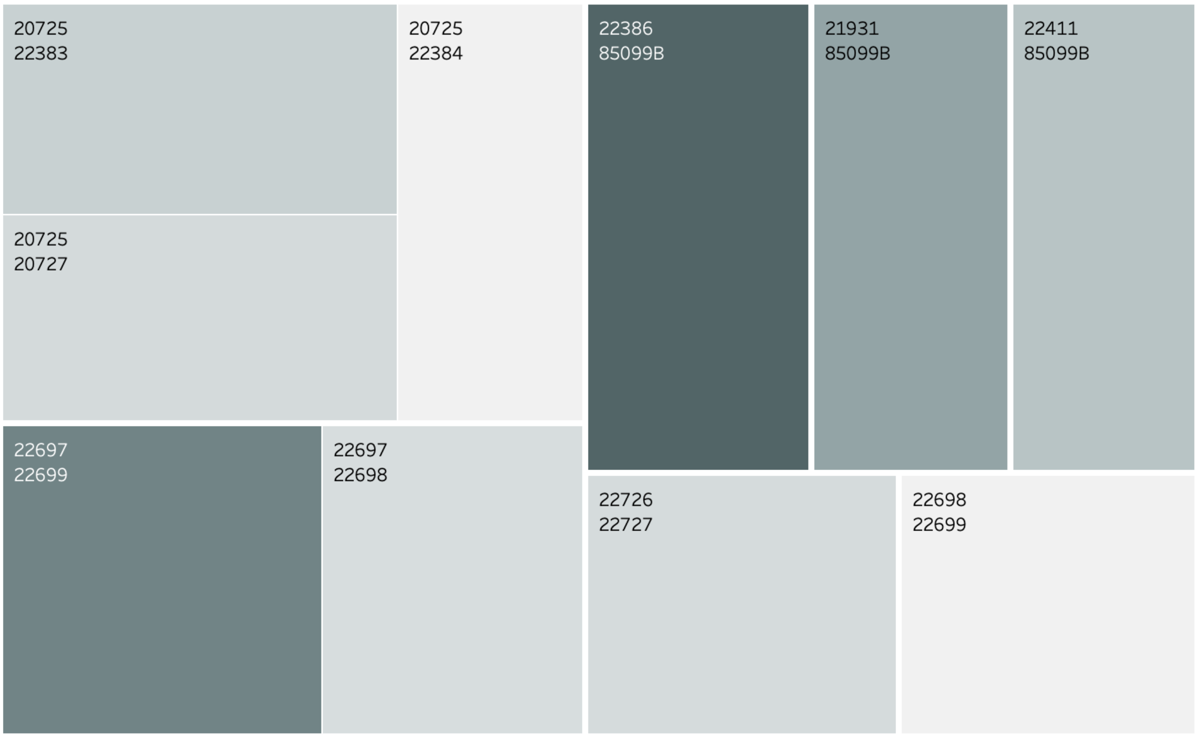
Implement loyalty rewards, VIP tiers, and exclusive benefits for customers who consistently make purchases past month 6.

This dataset only provides customer purchase behaviour, but a complete retention strategy requires tracking additional metrics.By integrating feedback mechanisms and behavioural tracking, businesses can proactively reduce churn and improve customer retention in the first 2-3 months. Without these additional insights, churn reduction efforts remain reactive rather than strategic.

Market Basket Analysis: Top 10 Frequently Bought-Together Products

Market Basket Analysis identifies product pairs that are frequently purchased together, helping businesses optimize product recommendations, bundling strategies, and cross-selling opportunities.

The code to SQL code for Market Basket Analysis can be found here



Key Insights from the Top 10 Product Pairs

Product “85099B” appears in 4 of the top 5 pairs, indicating it is a high-demand item that customers frequently purchase alongside other products.

Product Pair (22386, 85099B) was purchased together 833 times, making it the strongest association in the dataset.

Several sequentially numbered products (e.g., 22697 & 22699, 22698 & 22699, 22726 & 22727) suggest that certain product lines or collections are frequently purchased together.

Product “20725” appears in multiple pairs, meaning it has strong complementary relationships with different items (22383, 20727, and 22384).

The high co-occurrence count (above 600 purchases per pair) indicates strong buying patterns, making these product pairs ideal for bundling and recommendations.

Business Strategies Based on Market Basket Analysis

Product Bundling & Upselling

- Create combo deals for frequently bought-together items (e.g., “Buy 85099B + 22386 together and save 10%”).

- Offer discounts on complementary items (e.g., “Buy 20725 and get 50% off 20727”).
- Design pre-packaged gift sets with commonly purchased items (e.g., seasonal or themed bundles).

Optimize In-Store and Online Recommendations

- Feature related products on product pages to encourage cross-selling.
- Use personalized email campaigns (e.g., “Since you bought 22697, you may also like 22699”).

Inventory Management & Stock Optimization

- Ensure frequently bought-together items are well-stocked to prevent losing sales.
- Improve warehouse placement by storing commonly paired items together for faster fulfilment.
- Analyse seasonal trends to stock up on high-demand product pairs before peak periods.

By leveraging market basket insights, the business can increase average order value, enhance customer shopping experiences, and improve product visibility through strategic recommendations and promotions.

Conclusion

The analysis conducted in this case study revealed significant opportunities for improving customer retention, optimizing product performance, and increasing overall revenue. Through RFM segmentation, we identified high-value customers and at-risk segments that require targeted retention strategies. Cohort analysis highlighted the importance of early customer engagement, while market basket analysis provided insights into product bundling opportunities to drive higher sales.

Key takeaways from this study include:

Customer Retention & Buying Behaviour

- 27.1% of customers are at risk of churn, with early-stage retention being a critical challenge.
- Only 7.97% of customers are highly engaged (Champions), emphasizing the need for loyalty-building strategies.
- Over 50% of customers stop purchasing after 2-3 months, highlighting the importance of post-first-purchase engagement.

Revenue Trends & Seasonal Performance

- Revenue mirrors order trends, with November (Black Friday) driving the highest sales (£1.52M).
- February is the weakest month (£0.52M), indicating a post-holiday slump.
- December cancellations (41.1%) led to significant revenue loss, requiring better return management.

Product Performance & Market Basket Insights

- Products like “85099B” are highly co-purchased, making them strong candidates for bundling and upselling.
- Sequential product codes are commonly bought together, suggesting customers prefer curated product collections.
- Strong product associations (600+ co-occurrence counts) should inform personalized recommendations and inventory decisions.

4. Business Implications & Next Steps

- Retention efforts must focus on the first 30–60 days to prevent early churn.
- VIP programs and loyalty rewards should be strengthened for long-term customer retention.
- Seasonal marketing strategies should be expanded beyond Q4 to stabilize revenue across the year.
- Optimized product bundling and cross-selling strategies can drive higher order values and repeat purchases.

By implementing these recommendations, the business can move toward a more structured, data-driven decision-making process. Future efforts should focus on continuous data tracking, customer feedback analysis, and refining predictive analytics models to sustain long-term growth and profitability in the highly competitive e-commerce landscape.

The link to SQL code for overall project can be found here