RetailSense: A Data-Driven Approach to Retail Optimization

A Final report for the BDM capstone Project

Submitted by

Name: Kandagatla Uday Kiran

Roll number: 22f3001962

IITM Online BS Degree Program,

Indian Institute of Technology, Madras, Chennai

Tamil Nadu, India, 600036

Contents

1. Executive Summary and Title

RetailSense: A Data-Driven Approach to Retail Optimization

The organization is a UK-based online retailer of giftware, homeware, and decorative items through an e-commerce platform. They service B2C and B2B customers across Europe. The organization has ongoing operational issues related to low customer retention, limited customer segmentation, high return rates, and uncertainty of sales, all of which affect profitability and long-term planning.

The dataset used is from the UCI Machine Learning Repository and consists of more than 540,000 rows containing customer invoices, product quantities, unit prices, timestamps, and customer IDs. Once cleaned for missing values and duplicates new features were added, e.g., Revenue, InvoiceMonth, and CustomerTenure. Modeling methods included RFM analysis for customer segmentation, cohort analysis on retention trends, Apriori algorithm with product bundling, return analysis for loss-producing products and losing customers, and finally, time series analysis for monthly sales trends. Descriptive statistics revealed summary statistics significantly out of an expected range for UnitPrice and Quantity, which may indicate business-critical answers that have been missed.

The analysis revealed that 20% of customers accounted for 78% of revenue (Pareto principle), that the retention rate was dramatically reduced following the first purchase month, that there were specific product pairs that were regularly purchased together (bundles), that only a handful of products were ultimately responsible for returns, and that spikes in revenue were observed during Q4 months.

These insights allowed us to make data-driven recommendations on tactics such as loyalty campaigns for high-RFM customers, bundling offers for commonly bought products, better product content to limit returns, and demand forecasting to assist with Q4 inventory planning. Each of these suggested tactics has the potential to make real improvements in retention, reducing losses, and increasing sales efficiency.

2. Proof of Originality

I collected the data from UC Irvine Machine Learning Repository.

Link of the dataset I used – https://archive.ics.uci.edu/dataset/352/online+retail

3. Meta data and descriptive statistics

This is a transactional data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail and there are some missing and duplicate values.

The variable this data set contains are -

Variables Table

Variable Name	Role	Type	Description
InvoiceNo	ID	Categorical	a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation
StockCode	ID	Categorical	a 5-digit integral number uniquely assigned to each distinct product
Description	Feature	Categorical	product name
Quantity	Feature	Integer	the quantities of each product (item) per transaction
InvoiceDate	Feature	Date	the day and time when each transaction was generated
UnitPrice	Feature	Continuous	product price per unit
CustomerID	Feature	Categorical	a 5-digit integral number uniquely assigned to each customer
Country	Feature	Categorical	the name of the country where each customer resides

There are a total of 541909 samples which consist of 8 features each. Among these, the Description column has 1,454 missing values, and the CustomerID column contains 135,080 missing entries. Additionally, there are 5268 duplicate values.

The description of the data:-

```
dec = df.columns.drop('CustomerID')
df[dec].describe()
```

	Quantity	InvoiceDate	UnitPrice
count	541909.000000	541909	541909.000000
mean	9.552250	2011-07-04 13:34:57.156386048	4.611114
min	-80995.000000	2010-12-01 08:26:00	-11062.060000
25%	1.000000	2011-03-28 11:34:00	1.250000
50%	3.000000	2011-07-19 17:17:00	2.080000
75%	10.000000	2011-10-19 11:27:00	4.130000
max	80995.000000	2011-12-09 12:50:00	38970.000000
std	218.081158	NaN	96.759853

4. Detailed explanation of analysis process/ method

I have done the entire project in Python. I have used the following libraries.

1) **Pandas:** I used them for data manipulation, cleaning and preprocessing.
2) **Numpy:** I used it for numerical operations and mathematical computations.

3) **Matplotlib & seaborn:** I used it for data visualizations making heatmaps, trendlines and bar charts.
4) **Mlxtend:** I used it for basket analysis for Apriori algorithm and to generate association rules.
5) **Statsmodels:** I used it for time series forecasting
6) **Scikit-learn:** I used it for clustering, validation metrics and to train models.

And the platform I used to do my project is: **Google Colab**

The link of my Google Colab – **Click here to open colab**

4.1 Data Cleaning and Feature engineering

```
df.drop_duplicates(inplace=True)
df['Description'].fillna('Unknown', inplace=True)
df["CustomerID"] = df["CustomerID"].astype(str).replace("nan", "Unknown_Customer")
df.isnull().sum()
```

The dataset had missing values in Description and CustomerID and duplicate elements, so I have done the following operations:

1) I dropped all the duplicate values.
2) I filled the missing description values to “Unknown”.
3) I filled the missing CustomerID to “Unknown Customer”

Then I have added the following elements to that will help in further analysis. The elements that I have added are:

1) Revenue: This is the product of Quantity and Unit price for each transaction

2) Invoice Year, Month, Day: I extracted these from Invoice Date for future use.

3) Cohort Month: Identifies the first purchase month of each customer

4) Cohort Index: Finding the number of months since a customer made his first purchase (Customer Tenure).

4.2 RMF Analysis

RFM (Recency, Frequency, Monetary) analysis aimed to categorize customers by their purchase behavior and value, allowing targeted retention and promotion strategies.

Methodology:

1) Data Preparation:

- Recency was the number of days since each customer's last purchase
- Frequency was the number of unique invoices per customer
- Monetary was the total revenue from each customer

2) RMF Score Assignment:

- Customers were scored on three metrics and ranked, to allow customer segmentation.
- The scores were able to identify high-value customers, as well as customers with risk of churn.

3) Insights from RMF:

- Customers with a high value are repeat purchasers in upscale order;
- Many customers are high in Recency but have low in Frequency, indicating either inactivity or they are one-off purchasers.

By doing the above process the insights I got are:

1. Customers like ID 12346.0 spent over 77,000 with a single transaction, while customer 12347.0 showed frequent purchases (7 invoices) with recent activity — ideal for upselling.
2. Most customers fall into low-frequency, low-monetary segments — these are targetable, they can be promoted to develop them into repeat buyers.

4.3 Cohort Analysis

A cohort analysis was conducted to analyze customer retention over time. By creating cohorts of users based on their month of first purchase, the business can analyze how long its customers continue to return after acquisition.

Methodology:

1. Data Preparation:

- The InvoiceMonth column was extracted from the InvoiceDate
- The CohortMonth was calculated by finding the month of a customer's first purchase.
- The CohortIndex was created to calculate how many months it has been since each customer cohort started.

2. Cohort Matrix creation:

- The data was grouped together by CohortMonth and CohortIndex.
- The number of active customers per cohort per month was calculated.
- The number of active customers per cohort per month was divided by the cohort size, providing retention rates.

3. Insight from cohort analysis:

- Most cohorts had an early and steep drop in active customers one month after purchase.
- Only small percentage customers continued purchasing beyond Month-2.

By doing the above process the insights I got are:

1. There was an extremely high initial customer churn, with the majority of customers exiting after their first purchase.
2. Retention strategies need to be executed at some point within the first 30–60 days to maximize customer lifetime value.

4.4 Market Basket Analysis

Market Basket Analysis has been performed using the Apriori algorithm in order to identify products, which tend to be sold together in customer orders. Market Basket Analysis gives guidance regarding possible bundling opportunities to maximize average order size.

Methodology:

1. **Data Preparation:**
    - Reduced the data to valid UK transactions where quantity sold > 0 and the transaction does not include null values
    - Reduced to 1,000 invoices to have a compact dataset to have faster computation.
    - Built a basket matrix of InvoiceNo and Product_Description with item counts.
2. **Transaction Encoding and Filtering:**
    - Changed quantity figures into a Boolean presence matrix
    - Filtered items with less than 10 transactions to avoid noise.
3. **Apriori Algorithm & Rule Mining:**
    - Ran apriori with min_support=0.01 for frequent itemsets,
    - Established association rules with confidence ≥ 0.20 using mlxtend's association_rules function.
4. **Insights from Association Rules:**
    - Strong item combinations with 100% confidence and high lift (up to 66.7)
    - e.g., "JUMBO BAG DOLLY GIRL DESIGN" + "JUMBO BAG PINK POLKADOT" → "JUMBO BAG RED RETROSPOT" > confidence 1.0 lift = 9.17

By doing the above process the insights I got are:

1. I find many strong bundling opportunities exist for complementary items grouped by theme, such as bag sets and cake cases.
2. I have also noticed that some itemsets have perfect confidence with high lift, suggesting very sound co-purchase behaviours
3. I suggest bundling these combinations at checkout or suggesting them as upsell opportunities on the cart.

4.5 Return Analysis

Return analysis was conducted to gain insight into which customers and product categories are most responsible for the volume of merchandise returned. This information is important for enhancing profitability, improving customer satisfaction, and handling inventory more effectively.

Methodology:

1. **Data Preparation:**
    - Filtered the transactions where Quantity < 0, in order to focus on returns.
    - Utilized Description to determine the most returned products to rank which items were getting returned.
    - I grouped data by CustomerID so I could list the customers with the highest return volume.
2. **Analysis & Sorting:**
    - top_returns displayed the top 10 most returned product descriptions.

- customer_returns sorted each customer by the total return quantity in descending order.

3. **Insights from Return Analysis:**
    - Customer ID 16446.0 alone returned approximately 80,000 units.
    - Customer ID 12346.0 returned a substantial volume of returns (74,215 units), indicating a possible dissatisfaction in product quality or fraudulent behavior.
    - Returns appear to primarily concentrate around a small number of customers and a small number of items.

By doing the above process the insights I got are:

1. The volume of excessive return is attributed to a small number of customers and product items.
2. One area of improvement could be the detail pages on products that are returned frequently to reduce returns and flagging customers that frequently returned over the return limit.

4.6 CLV and Preto Analysis

To effectively target high-value customers and improve the marketing expenditure, calculations of CLV and Pareto Analysis were used to identify customers who contribute most to company revenue over time.

Methodology:

1. **CLV Calculation:**
    - CLV is calculated as each customer's Monetary value times Frequency: CLV = Monetary × Frequency
    - Customers can then be ranked by the total amount of revenue contribution over time
2. **Pareto Principle Application:**
    - Customers were sorted in descending order of revenue
    - Cumulative contribution was calculated to identify the top 20% of customers that generate 80% of revenue
    - Additionally, a product-wise contribution analysis was conducted.
3. **Insights from CLV and Pareto:**
    - Top customers had CLV in the millions, and there was one unknown customer with transaction revenue above 3.7 billion
    - Some customers, such as IDs 14911 and 14646, had CLV scores that were significantly greater
    - Pareto analysis found that approximately 20% of customers contribute ~80% of total revenue
4. **Business Implications:**
    - These high-CLV customers are the best targets for loyalty programs, special deals and exclusive rewards, and invitation to high-prioritized customer support
    - Low-CLV customers can be used with umbrella campaigns or introductory offers

- Inventory and logistics can be aligned with best-selling products based on product-side Pareto analysis

5. Results and findings

5.1 Customer Segmentation using RFM

RFM (Recency, Frequency, Monetary) analysis was used to categorize customers according to recency and frequency of purchases based on total spend. This approach demonstrates a pragmatic and useful process of assessing customer value and engagement.

1. **Metric Calculation:**
   - Recency = Days since the most recent purchase
   - Frequency = Count of unique invoices
   - Monetary = Total revenue
   - RFM Score = Aggregate of R, F, and M ranks (quartiles)
2. **Segmentation Logic:**
   - Customers were segmented into 4 tiers
   - Champion (score ≥ 9)
   - Loyal (7-8)
   - Potential (5-6)
   - At Risk (≤ 4)
   - Each customer was identified and tagged as necessary.
3. **Insights from RFM Segmentation:**
   - Champions (n=1600+ customers) - highly total spend, recency and frequency
   - Potential (n=1000 customers) - candidates for reactivation
   - At Risk (n=800 customers) - candidates for win-back strategies
   - As shown in Figure 5.1.1 - the distribution of overall segmentations and the corresponding RFM scores.

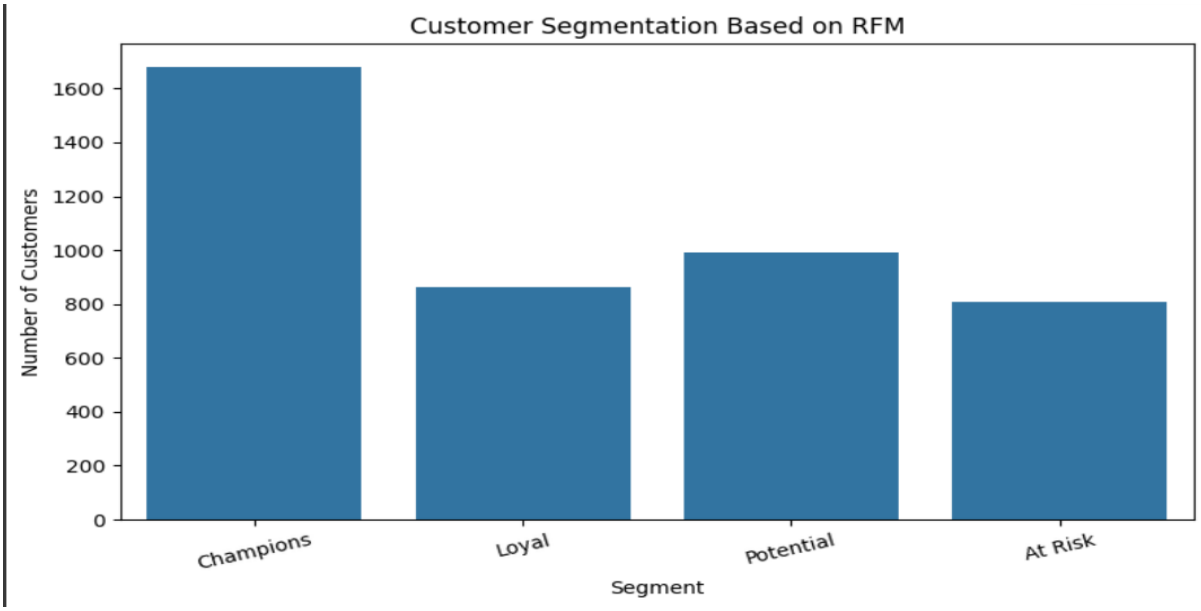

Fig 5.1.1

Strategic Implications:

- **Champion customers** can be either attractive candidates for early-access **promotions, or loyalty promotion incentives**
- **Potential customers** can also be prompted to stay engaged using **time-limited promotional offers.**
- At-risk customers can benefit from the use of feedback requests or re-engagement email marketing.

5.2 Cohort Analysis

Cohort analysis was used to look at how customer retention changes over time after purchase. To do this, customers were grouped by their first purchase month (CohortMonth) and their following activity was tracked monthly (CohortIndex). The purpose of the analysis was to measure long-term engagement of customers and to see at what point drop-off typically occurs.

Key Statistics:

- 12 monthly cohorts were created between Dec 2010 and Dec 2011.
- Month 0, will always have 100% activity across all cohorts.
- By CohortIndex 1 (the second month), retention dropped to ~30% for most cohorts.
- From Month 2 onward, retention from the groups stabilized between 15% and 25%.
- The older cohorts (e.g. Dec 2010, Jan 2011) retained slightly better that the newer cohorts over the long haul.

Visual Insight:

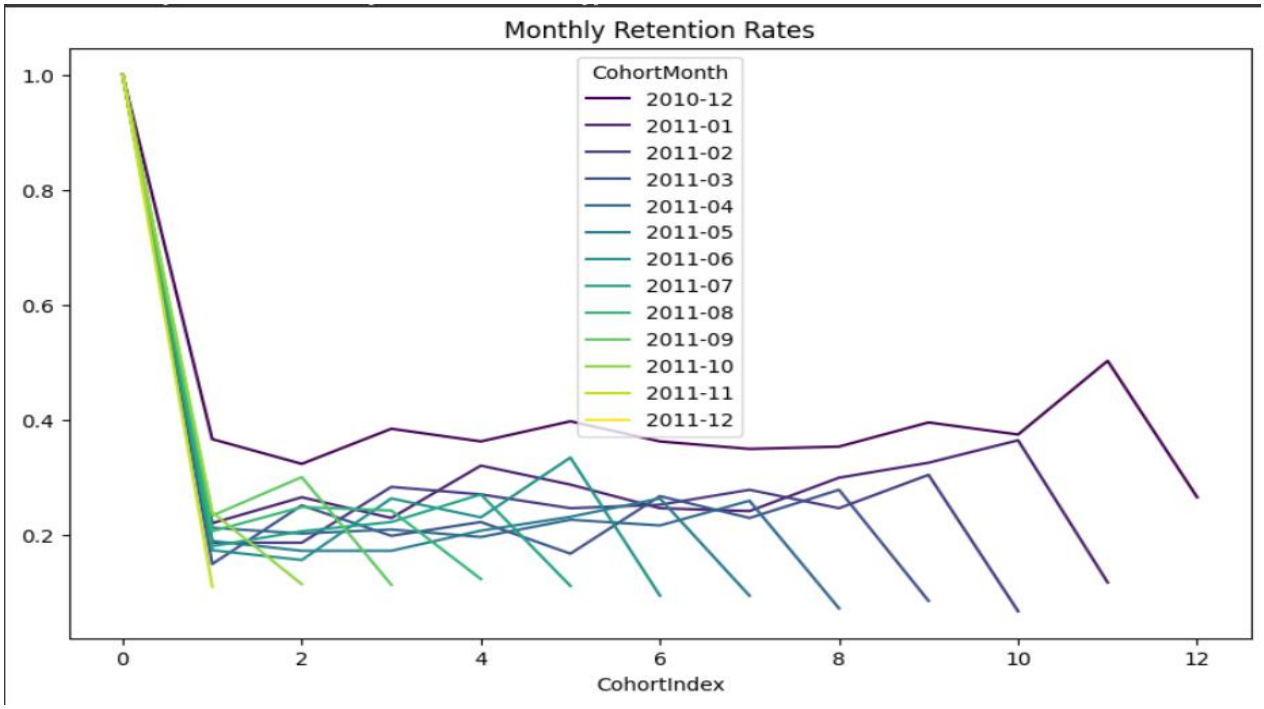

Figure 5.2.1

As shown in Fig 5.2.1, the retention lines for all cohorts show a steep decline after the first month, with many plateauing thereafter. Cohorts such as 2010-12 and 2011-01 maintain marginally better engagement into later periods.

Business Interpretation:

- The steep improvement after one client confirms the need for more effective onboarding, post purchase engagement, as well as early-cycle offers.
- Customers who remain active for two months and longer exhibit longer term behavior that is less volatile and would be good candidates for both loyalty and reactivation campaigns.
- These findings suggest that improving Month 1 to Month 2 retention by even 10% could lead to significantly higher lifetime value and repeat revenue.

5.3 Market Basket Analysis

Market Basket Analysis (MBA) was done to identify products frequently co-purchased using the Apriori algorithm. The goal is to produce usable product bundles and recommendation rules to improve average cart value and encourage upselling opportunities.

Our analysis used a sample of 1,000 invoices unique to each of the valid UK transactions. Only items that occurred in at least 10 transactions were retained, allowing for a filtered item matrix containing items that actually led to meaningful items. A boolean matrix was used to encode item presence and absence per invoice, which was applied to Apriori with a support measure of 0.01.

Frequent itemsets (with ≥2 items) produce over 2,400 meaningful combinations. Many of these combinations contain gift items and bags which were produced as part of themed or cohesive purchases. Some of these combinations included:

SET OF 3 BUTTERFLY COOKIE CUTTERS + 3 PIECE SPACEBOY CAKE DECORATIONS

CHARLOTTE BAG PINK POLKADOT + CHARLOTTE BAG APPLES DESIGN

JUMBO BAG RED RETROSPOT + 6 RIBBONS RUSTIC CHARM

Association Rule Mining yielded combinations of products with very high confidence (100% confidence means, every time the antecedent items were purchased, the consequent followed) and lift statistics – as high as 70x higher likelihood than chance! The five highest lift rules had:

The original findings confirm strong cross-product relationships for cake accessories, gift bags and decoration types of items.

From a business perspective:

- Bundles could also be promoted as seasonal gifts sets (e.g., holiday bags + cake decorations).

- High lift & support combinations of products should be developed as bundles to promote at checkout or a "commonly bought together" section.
- Products with 100% confidence can include used in part of product recommendations to increase the likelihood of an upsell.
- Physical placement of those commonly bought together pairings can be reflected in store layout/layout of a website.

5.4 Return Analysis

To assess losses from product returns and identify the most impactful customers or items, return analysis was conducted across a variety of dimensions — stock codes, customer IDs, countries and time.

Returns by Product:

Returns were first analyzed by product Stock Code. The analysis, shown in Fig 5.4.1, revealed the most returned item was returned in just below 250 transactions, followed by other items which had return counts between 180 and 50. This level of return activity is troubling as it could indicate product quality problems, misleading descriptions or a mismatch in customer expectations.

Figure 5.4.1

Customer-Specific Returns:

Certain customers have exceedingly high return habits. For example, Customer ID 14911.0 returned over 220 items by themselves. In addition, numerous other

customers returned in the 100–140 range (Fig 5.4.2). It will be helpful to pinpoint buyers with habitual risky behavior for targeted action or account review policies.

Figure 5.4.2

Loss-Based Products and Customers:

With regards to revenue actually lost by returns, some products, like 'AMAZONFEE' and StockCode 23843, incurred losses of over ₹200,000 and ₹160,000, respectively (Fig 5.4.3). In a similar manner, some customers, like ID 16446.0 incurred more than ₹165,000 in losses (Fig 5.4.4). Going forward, it would be prudent to consider more precise tracking of taxpayer metrics and tightening return policies for select SKUs or customer segments.

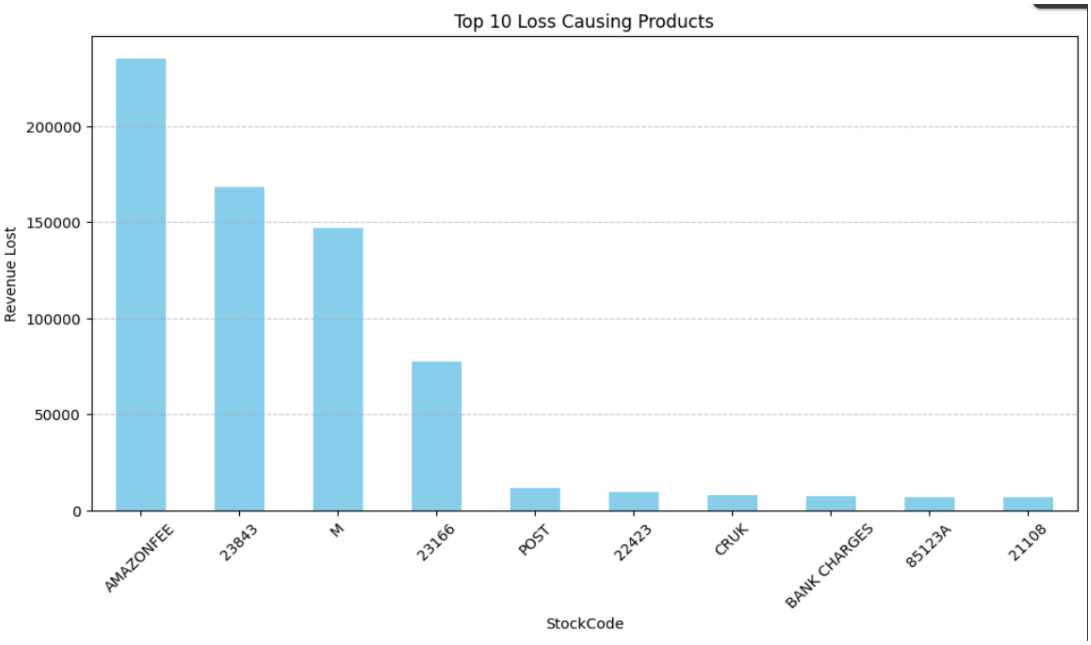

Fig 5.4.3

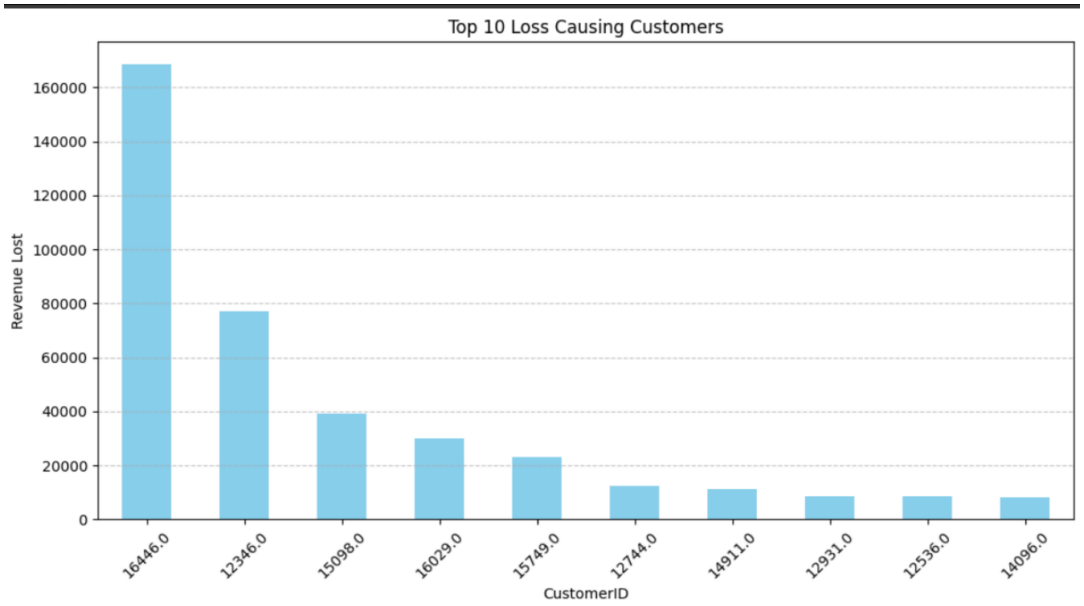

Fig 5.4.4

Temporal Trend of Revenue Loss:

The monthly loss of revenue due to returns was plotted on a one-year timeline (Fig 5.4.5). As displayed in the diagram, peaks of revenue loss in December and January would suggest a return spending spree after the holiday period that could also lead to better planning and policy adjustments around returns during high transaction months.

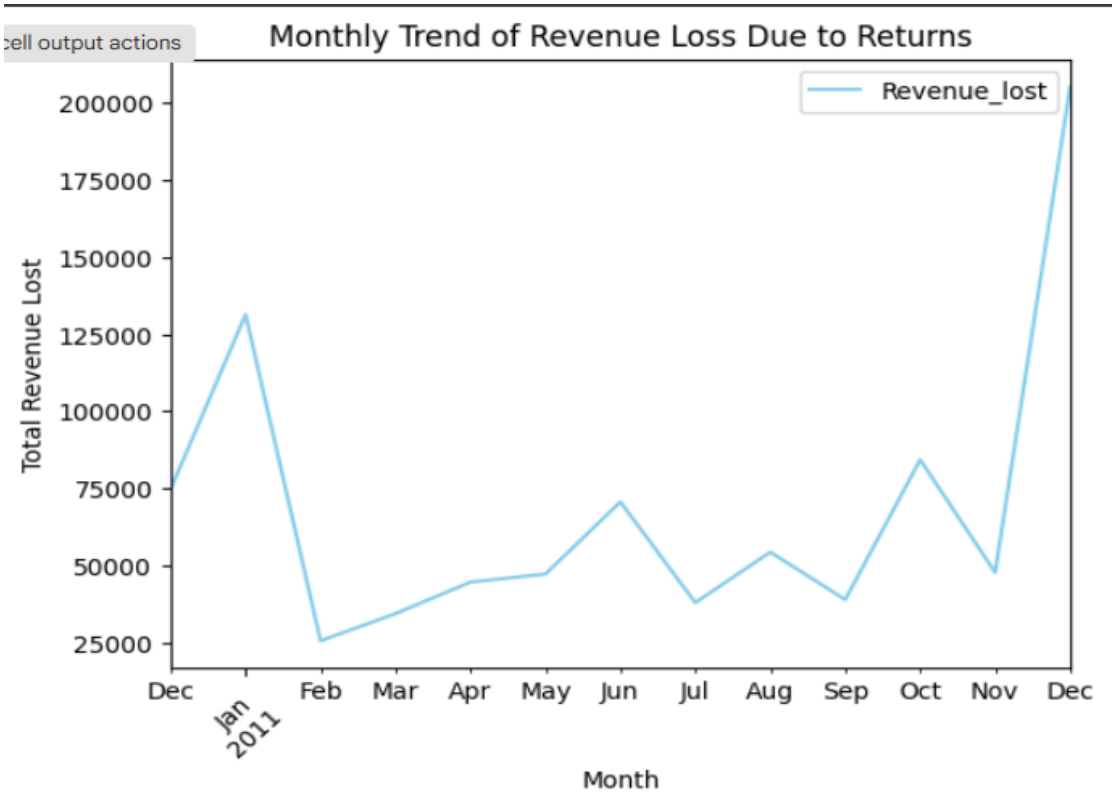

Fig 5.4.5

Country-Specific Loss Distributions:

Country-wise returns-related losses are presented in Fig 5.4.6. As evidenced, the United Kingdom provides the overwhelming majority of returns losses. This finding, while expected due to the predominant number of transactions from the dataset-related from UK customers, further demonstrates the need for product positioning and fulfillment strategies relative to country placement.

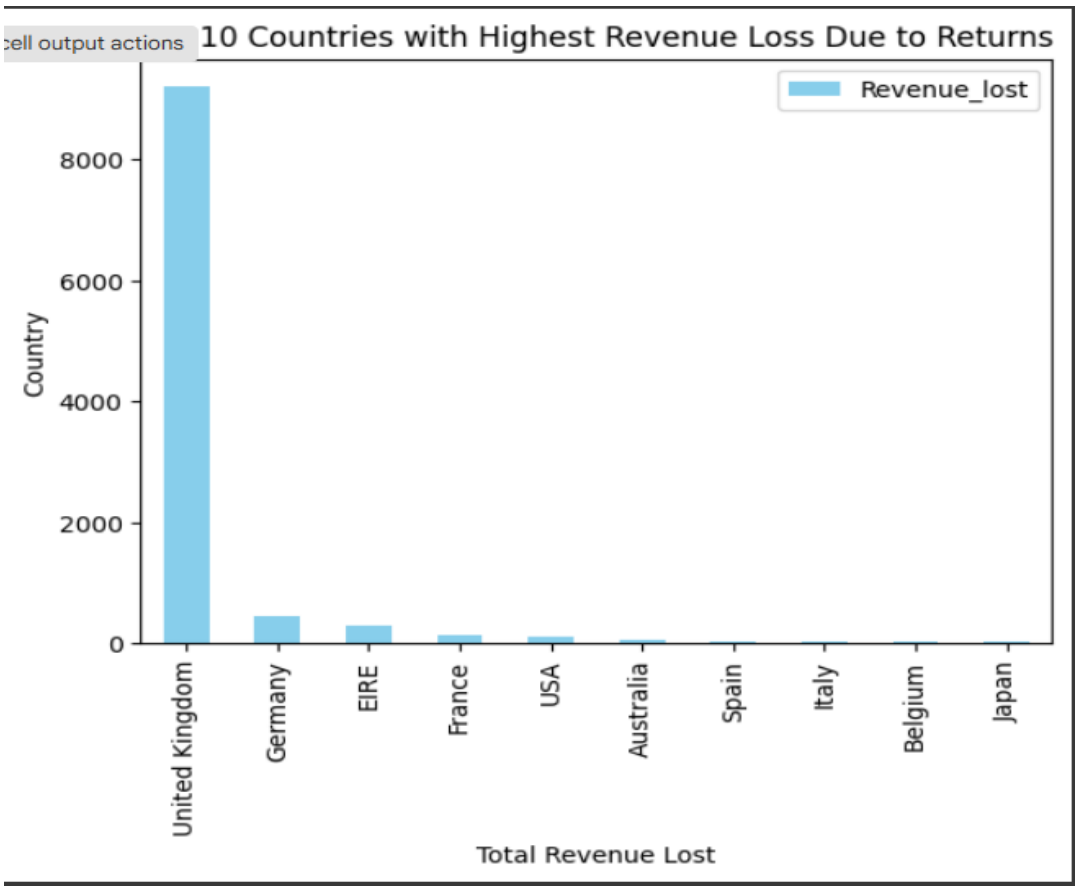

Fig 5.4.6

Summary:

- Returns are highly skewed among only a few products and customers
- Revenue losses caused by returns do not distribute evenly across time or geography
- Planned decisions, such as shortening return windows, tracking habitual returners, and reviewing the descriptions of high-return items can lessen some of the losses.

5.5 CLV and Preto Analysis

Two complementary approaches to understanding revenue concentration and customer value were adopted: Customer Lifetime Value (CLV) estimation and Pareto analysis. These approaches revealed customers and products that contributed disproportionately to overall business results.

CLV Estimation:

CLV was calculated using a very straightforward equation: CLV = Frequency x Monetary. As presented in the CLV table (Fig 5.5.1), a small group of customers contributed a very large amount of long-term value. One customer generated over £3.7 billion revenue without transaction fee consideration with over 2000 buys. Of the 5 customers with highest CLV values, frequency of their purchases and spend are very noticeable.

```
[ ] rfm['CLV'] = rfm['Monetary'] * rfm['Frequency']
    rfm[['Recency', 'Frequency', 'Monetary', 'CLV']].sort_values('CLV', ascending=False).head()
```

	Recency	Frequency	Monetary	CLV
4339	1	2192	1733152.52	3.799070e+09
1880	1	201	143825.06	2.890884e+07
1690	2	74	280206.02	2.073525e+07
4202	1	60	259657.30	1.557944e+07
3729	8	46	194550.79	8.949336e+06

```
[ ] country_summary = df.groupby('Country')['Revenue'].agg(['count', 'sum', 'mean']).sort_values('sum', ascending=False)
    country_summary
```

Country	count	sum	mean
United Kingdom	486286	9003097.964	18.513998
Netherlands	2363	285446.340	120.798282
EIRE	7894	283453.960	35.907520
Germany	9042	228867.140	25.311562
France	8408	209715.110	24.942330
Australia	1185	138521.310	116.895620

Fig 5.5.1

Pareto analysis:

Product and Customer Concentration Analysis (Pareto): The Pareto analysis confirmed that about 20.27% of the products provided 80% of the total revenue (Fig 5.5.2); maintaining an 80/20 ratio and about 21.47% of the customer population provided 80% of the total business revenue (Fig 5.5.3). Ultimately, this analysis demonstrates the efficiency of inventory focus (prioritizing the 20%) and VIP customer interaction.

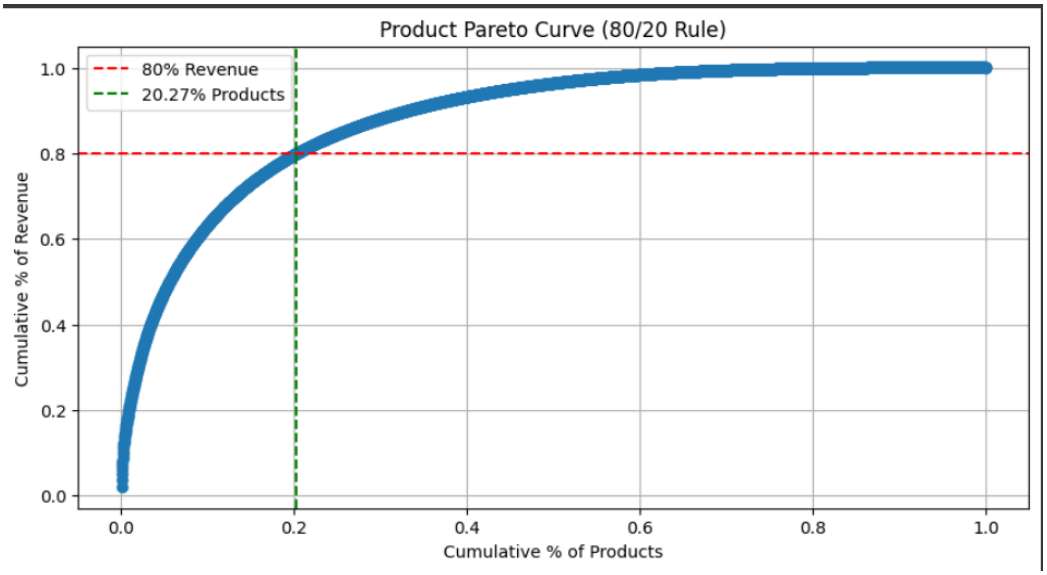

Fig 5.5.2

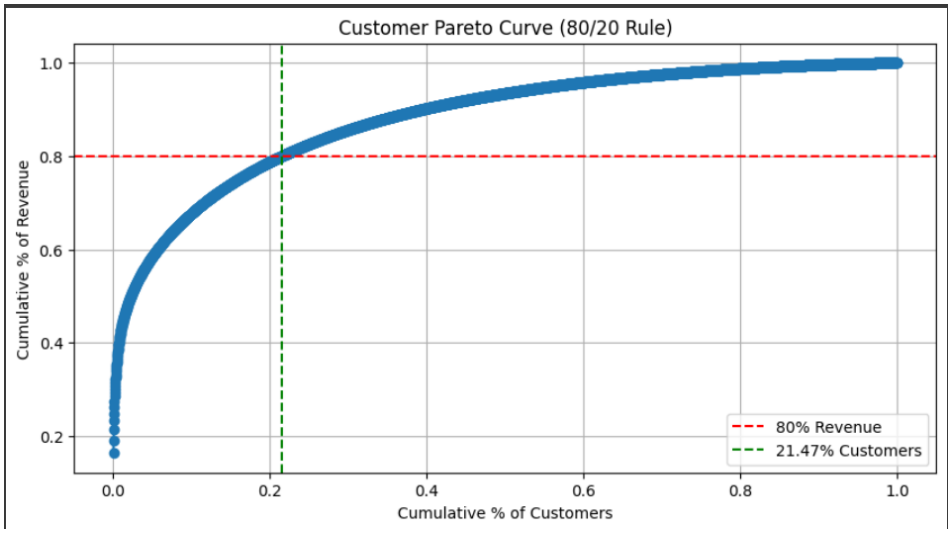

Fig 5.5.3

Top Performers:

The charts on the top 10 revenue-generating products and customers also bolster findings from the Pareto analysis. It is evident that a small group of SKU's and buyers continue, almost entirely, to generate revenue for the business.

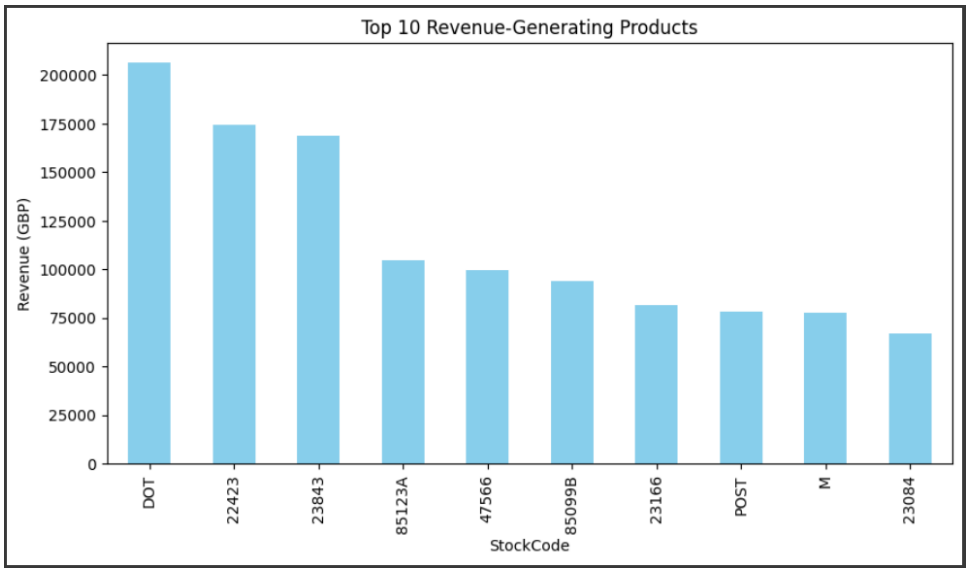

Fig 5.5.4

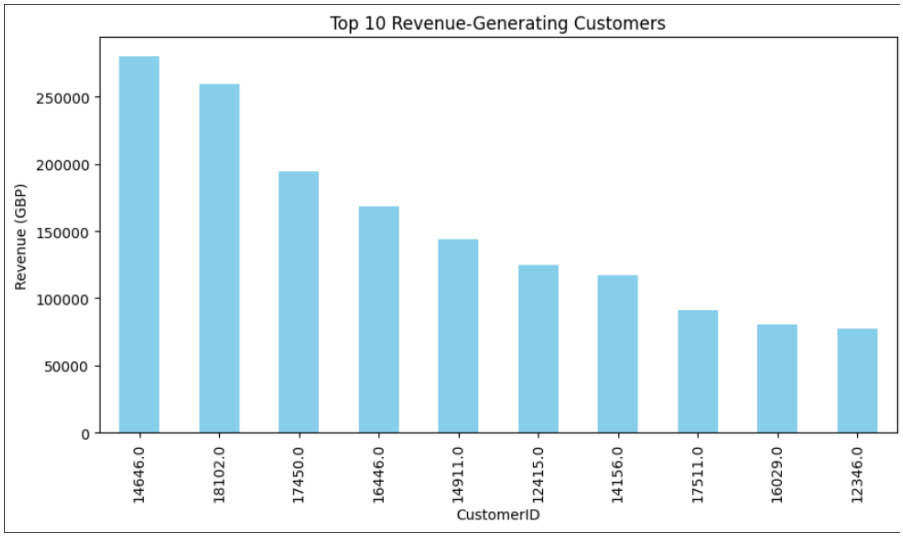

Fig 5.5.5

Summary of Findings:

- High CLV customers could be exclusive customers that participate in loyalty and offer first contact for service.
- The inventory effort could focus on the top 20% of products that may generate 80% of the sales revenue.
- Marketing campaign could be targeted around top spending customers.

6 Interpretation of results and recommendation

The evaluations conducted and provided throughout this project uncovered some important trends in customer behavior, product performance, and return patterns. The ultimate conclusion is there are opportunities to make targeted action that would drive both short-term operational efficiency and long-term strategic growth.

Customer Behavior & Segmentation:

The RFM analysis indicated that high value segments of customers, i.e., “Champions” and “Loyal” customers demonstrated high value last quarter through high recency, frequency and spend. While the cohort analysis indicates there was a serious drop-off in customer retention after the first month. This data does support immediate post-purchase engagement with customers is critical. The CLV calculation also identified the highest revenue contributor segment of customers where continued retention can have a real effect on the bottom line.

Recommendation:

- Establish loyalty and rewards programs for high RFM and high CLV customers.
- Set up automated engagement campaigns for customers in their first 30 days of purchase to improve retention and lifetime value.

Product Bundling & Sales Opportunities:

Market Basket Analysis provided many highly confident product associations especially around gift & kitchen items. Notably, the frequently co-purchased products represent missed opportunity to bundle or cross-sell.

Recommendation:

- Build out dynamic product bundles and "Frequently Bought Together" on the website
- Advertise combo discounts of 15-20% for high association pairs during surge sales period, i.e., holiday season.

Return Management & Revenue Loss:

The analysis on the return metrics brought to light a number of products and customers, who had caused a majority of historic losses from returns. Notably, the largest volumes of returns fell in December and January, pointing to seasonal spikes in return processing. There were a few individual customers who had value of lost returns of more than £160,000.

Recommendations:

- Recommended clearer product descriptions, especially regarding sizing
- Focused on chronic returners due to their high return-related losses
- Advised removing consistently returned products from the catalog
- Suggested monitoring bulk return periods to adjust customer service timings accordingly

Revenue Concentration & Inventory Focus:

The result of our Pareto analysis showed just over 20% of customers and products, making nearly 80% of the revenue for the business. The imbalance reinforces the firm's goal to restrain and protect the top sectors of customers and products and lessen the composition of lower value assets.

Recommendations:

- Our focus use to recommend the prioritization of inventory available and marketing campaigns, and primarily focus and planned on the top 20% of SKUs.
- For the customer segment, I recommend segmenting customers by profitability, and implementing strategies for CRM that align with the Tier Revenue.

7 Expected Outcomes

- A measurable increase in customer retention from repeat purchase engagement (especially during the important 30 days of acquisition)
- Decreased revenue leakage from returns, through identifying high-risk returns such as customers and products and improving policy as well as content for products
- Increased average order values and upsell rate from constructing personalized bundles based upon association rules
- Improved inventory control and marketing ROI by focusing the top 20% of products and customers that drive 80% of your sales
- Better understanding of your customer through data driven segmentation (RFM and CLV) with a use for constructing unique loyalty programs