Optimizing Online Retail Sales Using Data-Driven Insights

A Final report for the BDM capstone Project

Submitted by
Name: Mohit Tewari
Roll number:23f1002364

IITM Online BS Degree Program,
Indian Institute of Technology, Madras, Chennai
Tamil Nadu, India, 600036

Contents

1. Executive Summary

The UK-based online retail company (unknown name) specializes in selling unique homeware and gift items to both individual customers (B2C) and wholesalers (B2B). The company operates globally, with a diverse customer base purchasing through its e-commerce platform.

The online retail business faces challenges related to in demand forecasting, product performance evaluation, and return management. Unpredictable sales trends lead to inventory issues, while a significant portion of revenue comes from a small percentage of products. Additionally, frequent product returns impact profitability, requiring a deeper understanding of return patterns.

To address these challenges, a **data-driven approach** will be employed. Time-series forecasting models will be used to predict demand, helping optimize stock levels. Pareto analysis will identify high-performing products, enabling better sales and marketing strategies. Lastly, return analysis will highlight common return and problematic products, leading to improved quality control and reduced losses. In addition, Customer segmentation will be done for targeted marketing campaigns and loyalty programs. By leveraging these insights, the company can enhance operational efficiency, minimize costs, and improve customer satisfaction.

2. Proof of Originality

The data for the project has been collected from the UCI Machine Learning Repository.
Link of the dataset: https://archive.ics.uci.edu/dataset/352/online+retail

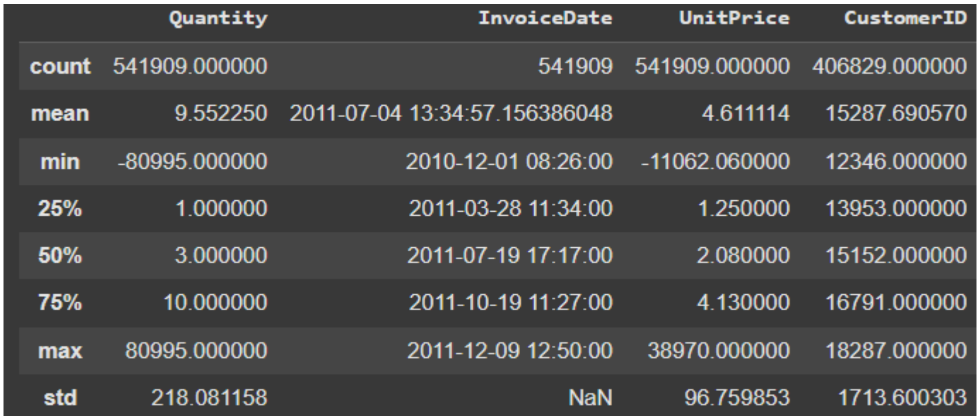
3. Meta data and Descriptive Statistics

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. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

- The Dataset contains 8 columns:
  - InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
  - StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
  - Description: Product (item) name. Nominal.
  - Quantity: The quantities of each product (item) per transaction. Numeric.
  - InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
  - UnitPrice: Unit price. Numeric, Product price per unit in sterling (GBP).
  - CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
  - Country: Country name. Nominal, the name of the country where each customer resides.There are missing values in Description and CustomerID.
- The dataset contains 541,909 samples.
- Missing Values:

    - Description: 1454 missing values.
    - CustomerID: 135080 missing values (about 24.9% of the sample).
- Duplicates: 5268 duplicate rows.
- Quick Summery Statistics:

  We can ignore CustomerID here.



	Quantity	InvoiceDate	UnitPrice	CustomerID
count	541909.000000	541909	541909.000000	406829.000000
mean	9.552250	2011-07-04 13:34:57.156386048	4.611114	15287.690570
min	-80995.000000	2010-12-01 08:26:00	-11062.060000	12346.000000
25%	1.000000	2011-03-28 11:34:00	1.250000	13953.000000
50%	3.000000	2011-07-19 17:17:00	2.080000	15152.000000
75%	10.000000	2011-10-19 11:27:00	4.130000	16791.000000
max	80995.000000	2011-12-09 12:50:00	38970.000000	18287.000000
std	218.081158	NaN	96.759853	1713.600303

Fig. 3.1 "Summery Statistics"

4. Detailed Explaination of Analysis Process/Method

- **Tools used for Analysis:**
    - Python and Python libraries:
        - Numpy
        - Pandas
        - Matplotlib, Seaborn
        - Statsmodel library etc
    - Google Colab:
        - All the analysis and code are stored here.
        - **Link of the Colab notebook**: https://colab.research.google.com/drive/1hj0biI58U_m53DxJO14TJvhFX6ARL0hP?usp=sharing

1. **Data Preprocessing and Cleaning**:

   The dataset included missing values in 'Description' and 'CustomerID' columns that needed imputation and there were over 5000 duplicate rows that needed removal. The following methods were used for resolving the issues:

    - Missing Description values were filled with 'Unknown'.
    - Missing CustomerID values were filled with 'Unknown_Customer'
    - Duplicated rows were removed.

   For the future analysis, some columns like revenue for revenue related analysis and day, month and year etc for trend analysis were needed. The following strategies were used to derive them:

    - New Features:
        - Revenue: Derived as Quantity * UnitPrice
        - Date based features like year, month, day, weekday etc were derived from 'InvoiceDate' column.

2. Pareto Analysis

The goal of Pareto analysis in this project was to identify the key products and customers contributing the most to total revenue and returns. This follows the **80/20 rule**, which states that a small percentage of products or customers often generate a large percentage of the revenue.

Methodology:

1. **Data Preparation**:
   - Extracted total sales revenue per product and per customer.
   - Sorted products/customers in descending order based on total revenue.
2. **Pareto Principle (80/20 Rule) Application**:
   - Calculated cumulative revenue contribution.
   - Identified the percentage of top products/customers contributing to 80% of total revenue.
3. **Insights from Pareto Analysis**:
   - A small percentage of top-selling products contributed to the majority of total revenue.
   - A small percentage of high-value customers were responsible for the majority of purchases.
4. **Visual Representation**:
   - Plotted Pareto chart to show cumulative revenue contribution and determine the cut-off for 80% contribution.
   - Created separate charts for sales contribution.

The above analytical approach provided us useful insights:

1. Top 20% of products contributed to nearly 80% of total revenue. Therefore, inventory efforts can be focused on them.
2. Top 20% of customers accounted for most of the purchases. Prioritizing marketing and loyalty programs toward them might be good.

3. Demand Forecasting

The aim of demand forecasting in this project was to predict future sales trends based on historical data. This would help in optimizing inventory management, improving stock availability, and minimizing overstock or stockouts

Methodology:

- **Data Preparation:**
  - Extracted sales data over time.
  - Aggregated total sales at a daily and monthly level.
- **Exploratory Data Analysis (EDA):**
  - Plotted **time series trends** to check for patterns like seasonality, trends, and cyclic behaviour.
  - Used Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots to determine p, d and q components of ARIMA model.
    - **Autocorrelation Function (ACF):**
      - ACF measures how current values are related to past values (lags).
      - If sales data is seasonal, ACF will show peaks at regular intervals (e.g., every 12 months for yearly seasonality).
      - If sales have a strong trend, ACF will slowly decrease over lags instead of dropping suddenly.

      - **Partial Autocorrelation Function (PACF):**
        - PACF removes indirect correlations between time points.
        - If PACF has a sharp drop after a few lags, it tells us how many past values (lags) we should use in our forecasting model.
        - If PACF shows high values at seasonal lags (like lag 12 for monthly data), we likely have seasonality.
    - Checked for stationarity using the Augmented Dickey-Fuller (ADF) test.
      - The **Augmented Dickey-Fuller (ADF) test** is a statistical test used to determine whether a time series is stationary or contains a unit root, meaning it is not stationary. Essentially, it checks if a time series is likely to follow a random walk and not have a constant mean and variance over time.
    - Stationarity was found and needed removal as models like ARIMA assumes non-stationarity in the data.
    - Applied first-order differencing (subtracting the previous month's revenue from the current month) to remove the trend and make the series stationary.
  - **Model and Parameter Selection:**
    - Chose ARIMA after analyzing the Autocorrelation Function (ACF), Partial Autocorrelation Function (PACF) plots and Augmented Dickey-Fuller (ADF) test.
    - p' was chosen as 0 because PACF had no significant lags; AR (0).
    - 'd' was chosen as 1 because ADF Test showed non-stationarity, but became stationary after first-order differencing.
    - 'q' was chosen as 0 because ACF had no significant lags; MA (0).
    - Both p and q values suggest no auto-regressor modal or moving average model will provide significant forecasting, yet I tried ARIMA with above p,d and q values.
- **Model Training & Evaluation:**
    - Data was split into training and test sets.
    - Chose ARIMA parameters carefully to minimize error.
    - Evaluated the model using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

4. Return Analysis

As product return can lead to large revenue losses, the goal of Return Analysis was to figure out which products are being returned or customers are returning frequently or generating huge revenue losses. In addition, identifying if return volumes follow a seasonal trend. This would help providing useful insight in return patterns to tackle or better manage revenue losses.

Methodology:

1. **Data Preparation (Filtering Returns)**
    - The dataset was filtered to extract returned transactions
    - A new column Revenue_lost was created to calculate losses from returns.
2. **Analysing Returns**
    - Top Products with High Return Counts were identified by counting how many times each product was retuned and sorted them in descending order.
    - Top Customers with High Return Counts were identified by counting how many times each Customer retuned an item and sorted them descending order.
    - Top Products with High Revenue losses were identified by summing up revenue loss for each product and sorted them descending order.
    - Top Customer who generated High Revenue losses were identified by summing up revenue loss for each product they returned and sorted them descending order.
    - Monthly trend of revenue loss due to returns was identified by plotting a line chart.
    - Top 10 countries with highest revenue loss due to returns was identified by plotting a bar char chart.

5. Customer Segmentation

The goal of customer segmentation was to identify distinct customer groups based on their purchasing behaviour. This would allow for targeted marketing strategies, personalized offers, and optimized customer engagement.

Methodology:

1. Data Preparation

Features Used for Clustering

To segment customers, we selected the following key features:

- **Total Revenue**: Total spending of the customer.
- **Total Transactions**: Number of purchases made by the customer.
- **Total Products Purchased**: Total quantity of products bought.
- **Unique Products Purchased**: Number of different products purchased.
- **Return Frequency**: How often a customer returned products.

Preprocessing Steps:

- Outlier ‘Unknown_Customer’ was removed due to past errors in clustering.
- Normalization, using MinMaxScalar, the data to ensure all variables were on a comparable scale**.**

2. Clustering Methodology

Applied K-Means clustering to segment customers based on their spending and purchase frequency.

For determining the Optimal Number of Clusters (k), two techniques were used:

1. Elbow Method:
    - Plotted inertia (sum of squared distances to cluster centres) for different values of k.
    - The elbow points suggested k = 2 or k = 3.
2. Silhouette Score:
    - Measured the quality of clustering for different values of k.
    - The highest silhouette score was at k = 2, confirming 2 distinct clusters.

3. Results

Two major customer segments were identified:

- Regular Shoppers (low spending, low returns, occasional buyers).
- High-Value Customers (frequent purchases, high spending, high returns).

5. Results and Findings:

1. Pareto Analysis (80-20 rule)

Analysis included identifying:

- Top 20% Products contributing to 80% revenue.

- Top 20% Customers generating 80% revenue.

1. **Product Pareto Analysis**:

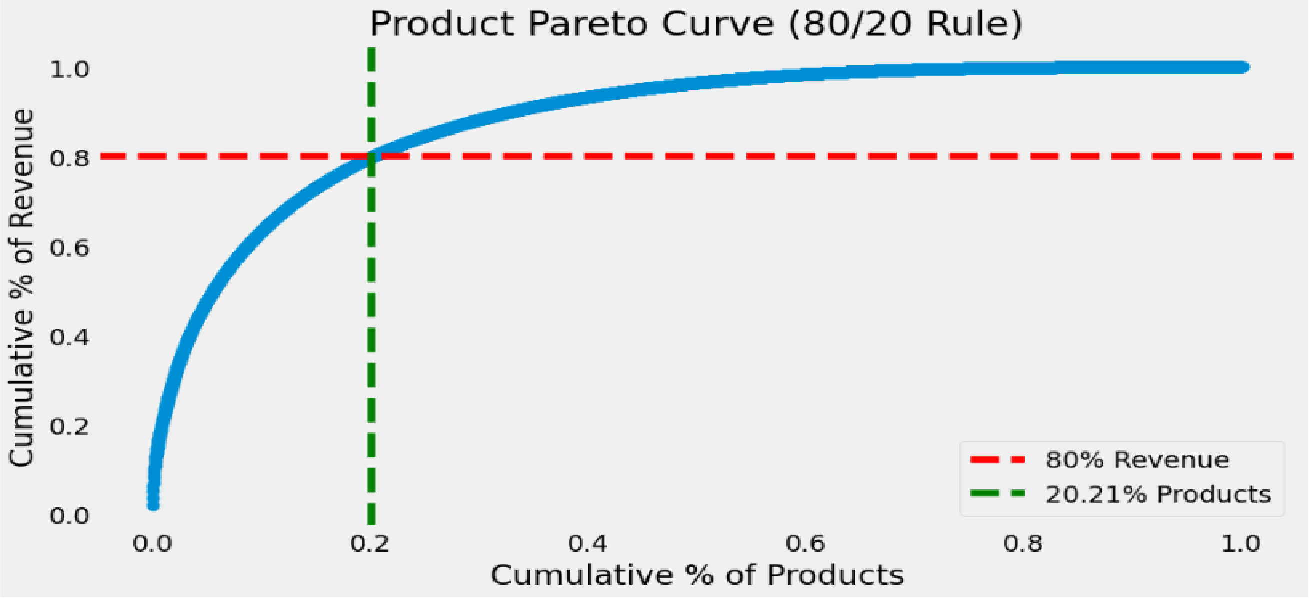


Fig 5.1.1 Product Pareto Curve

Interpretations:

a. Total Unique Products: **4,078**

b. Products contributing to 80% revenue: **824 (20.21%)**

c. Top Best-Selling Products:

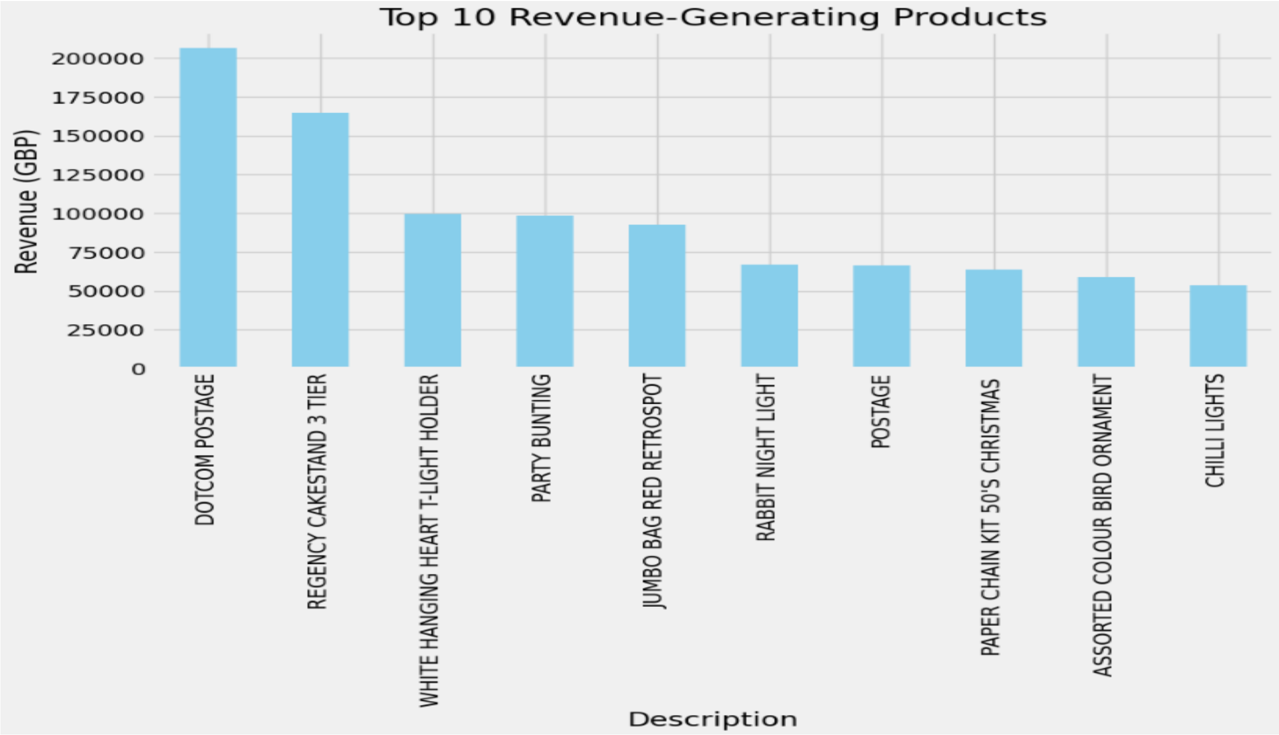


Fig. 5.1.2 Top 10 Revenue Generating Products

d. **Key Takeaway**: A small fraction of products generates the majority of sales, so inventory and marketing efforts should prioritize these high-revenue items.

2. **Customer Pareto Analysis:**

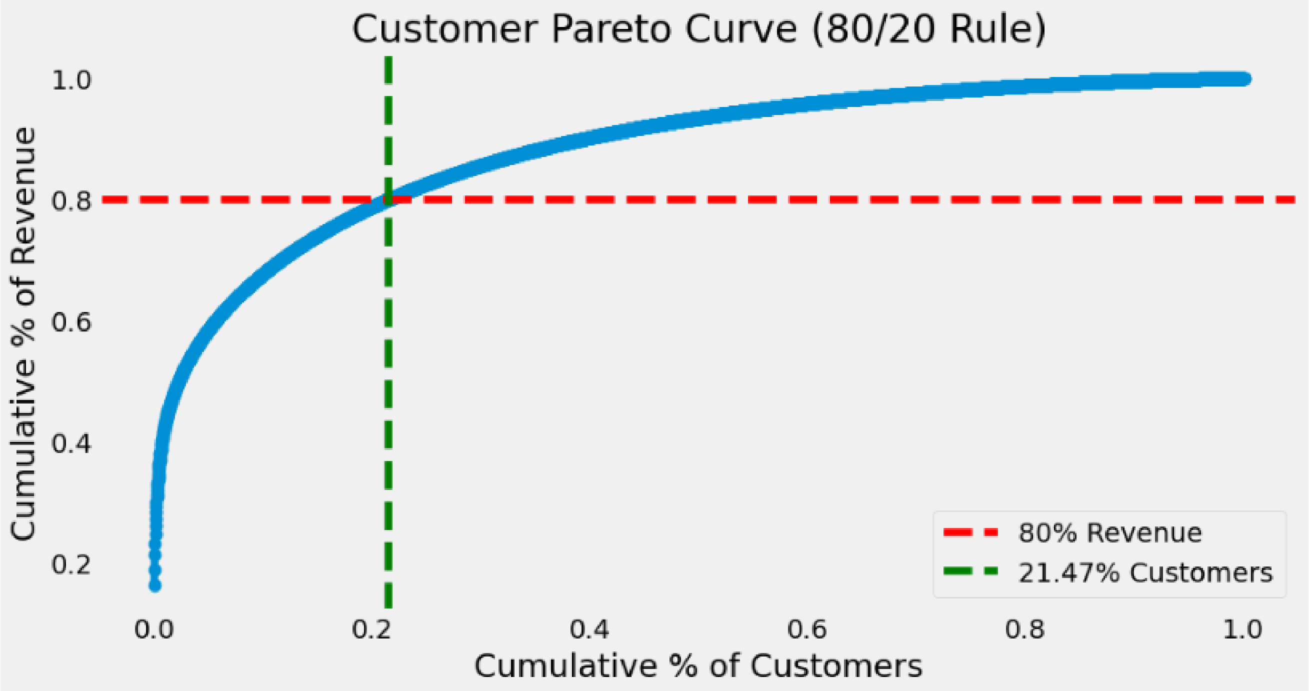


Fig 5.1.3 Customer Pareto Curve

Interpretations:

a. Total Unique Customers: **4,340**

b. Customers contributing to 80% revenue: **932 (21.47%)**

c. Issue Identified "Unknown_Customer":

   i. The largest revenue-generating "customer" is actually Unknown_Customer (missing CustomerID) with revenue over GBP 1.7M.

   ii. This could indicate guest checkouts, bulk B2B transactions, or data entry issues.

d. Top 10 Revenue-Generating Customers (Without ‘Unknown_Customer’):

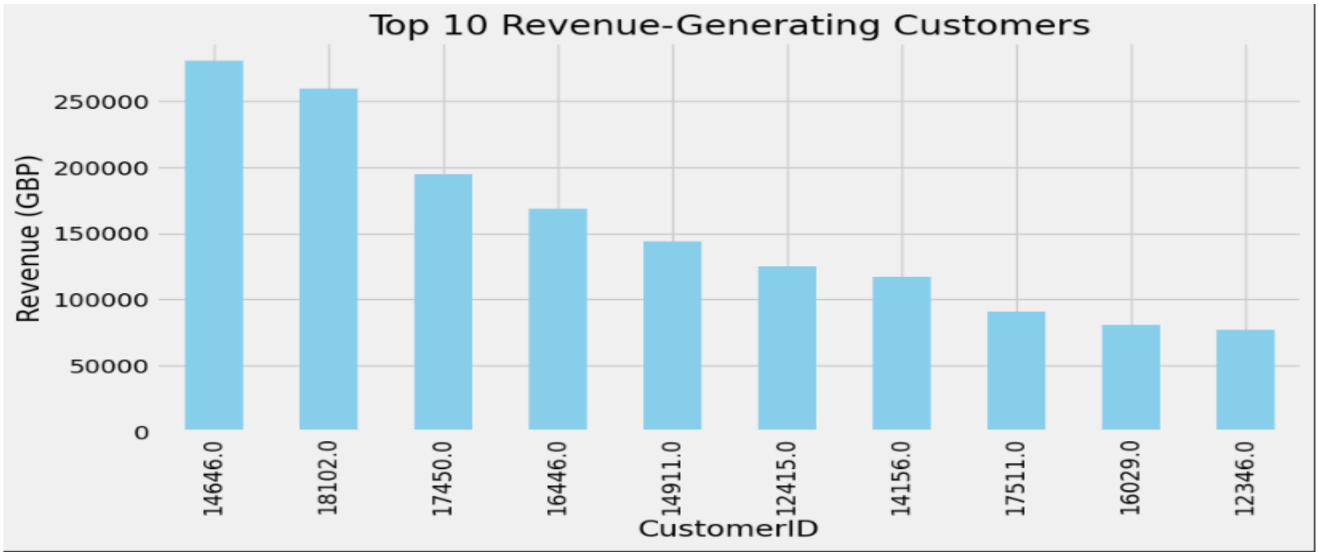


Fig. 5.1.4 Top 10 Revenue-Generating Customers

e. **Key Takeaway**: A small group of high-value customers drive most revenue, so personalized marketing and loyalty programs can maximize retention.

Summary:

- We found top contributing Products so inventory efforts can be focused on fewer products.

- We found top revenue generating Customers so we can prioritise marketing efforts and loyalty programs toward them.

2. Demand Forecasting:

1. Daily revenue Trend plot:

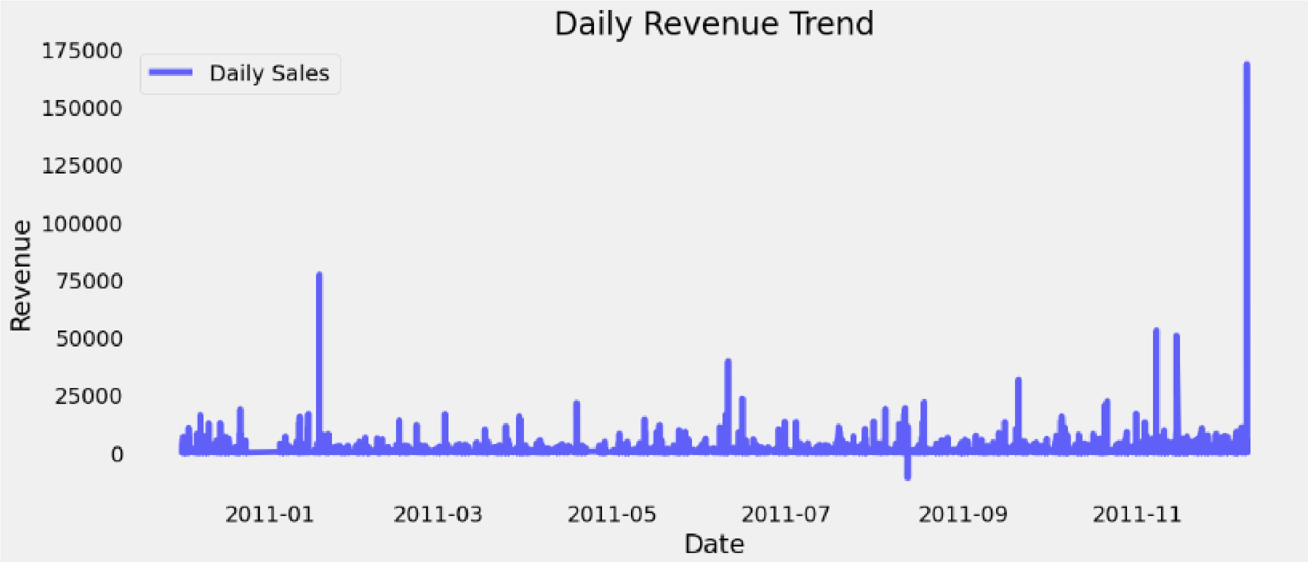


Fig 5.2.1 Daily Revenue Trend

Interpretations:

- High Variability - There are frequent spikes in revenue, suggesting irregular demand patterns.
- Some Extreme Outliers - Few days show huge sales peaks, likely due to bulk B2B orders or special promotions.
- Seasonality & Dips - Some days have very low sales, which could be weekends, holidays, or operational issues.

2. Monthly revenue Trend plot:

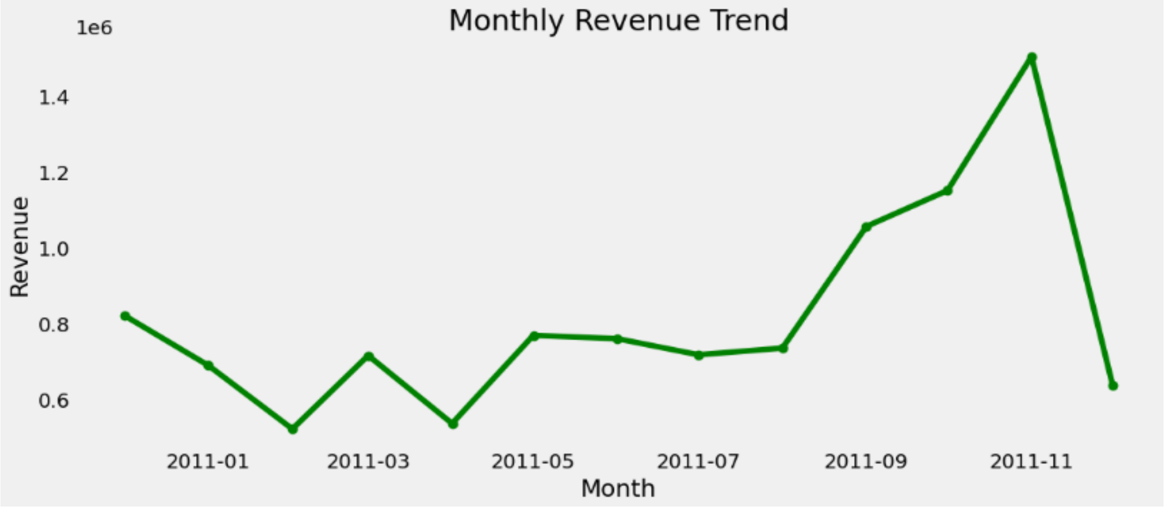


Fig 5.2.2 Monthly Revenue Trend

Interpretations:

- Cyclical Pattern - Revenue drops early in the year, rises later, showing possible seasonality.

- End-of-Year Surge - Sales peak in November-December, likely due to the holiday season.
- Sharp Decline in December - This could be end-of-year returns, reduced orders post-holidays, or missing data.

3. **Checking for Stationarity (Augmented Dickey-Fuller (ADF) test):**
   - ADF Statistic: -2.16
   - P-value: 0.22
   - Critical Values:
     - 1%: -4.12
     - 5%: -3.15
     - 10%: -2.71

   p-value (0.2196) is greater than 0.05, meaning that the time series is non-stationary. This suggests that it has trends or seasonality that we need to remove before applying forecasting models like ARIMA.
4. **After Differencing the Data:**
   - ADF Statistic after Differencing: -3.68
   - P-value after Differencing: 0.004
   - Critical Values after Differencing:
     - 1%: -4.22
     - 5%: -3.18
     - 10%: -2.73

   p-value is smaller than 0.05, meaning the time series is now stationary.
5. **Autocorrelation (ACF) and Partial Autocorrelation (PACF):**
   1. **ACF Plot:**

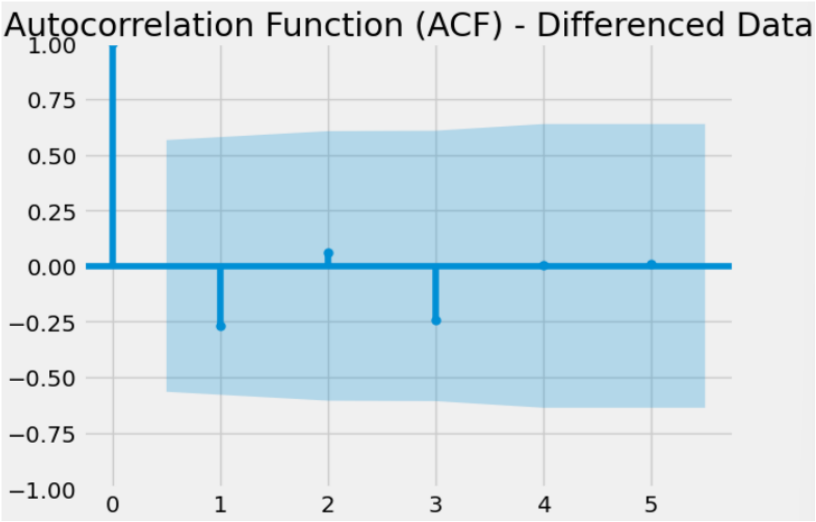


Fig 5.2.3 ACF Plot

Interpretations:

All the lag values come under the shaded region. No significant lags.

   2. **PACF Plot:**

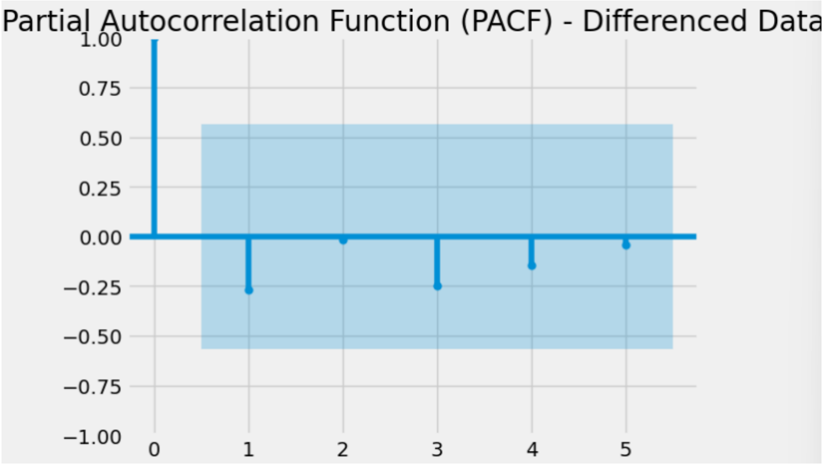


Fig 5.2.4 PACF Plot

Interpretations:

All the lag values come under the shaded region. No significant lags.

Conclusion: Forecasting will be ineffective, and using ARIMA or other time-series models won't yield good results.

Yet I tried to fit ARIMA.

6. Fitting ARIMA model:

- **Parameters for ARIMA:**
  - **'p' = 0**; PACF had no significant lags; AR (0).
  - **'d' = 1;** ADF Test showed non-stationarity, but became stationary after first-order differencing.
  - **'q' = 0;** because ACF had no significant lags; MA (0).

- **Forecast vs Actual Plot:**

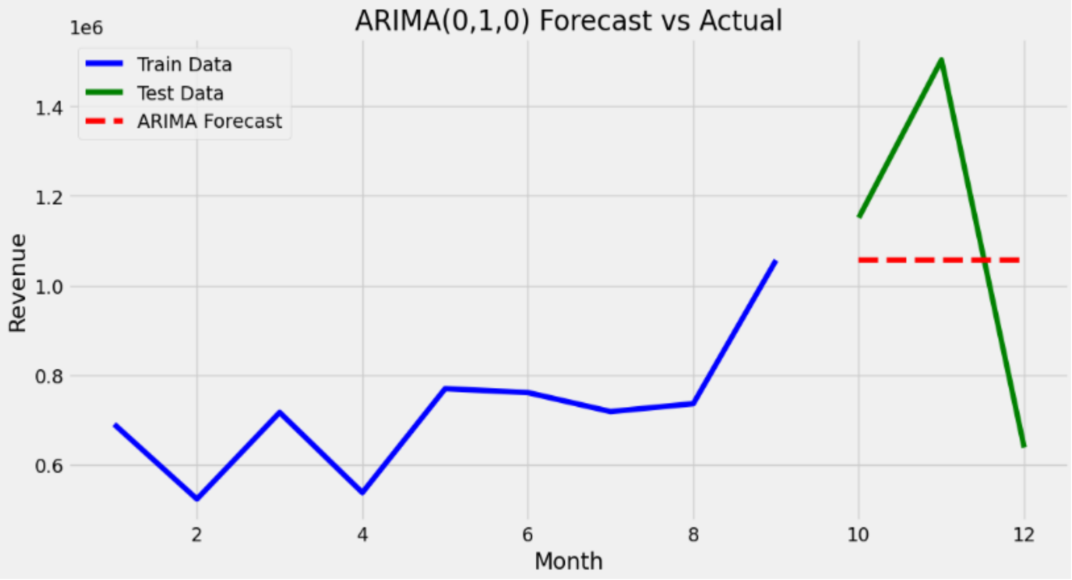


Fig 5.2.5 Forecast vs Actual

- **Error terms:**
  - Mean Absolute Error (MAE): 320301.66
  - Mean Squared Error (MSE): 128150332679.47
  - Root Mean Squared Error (RMSE): 357980.91

Interpretations:

1. **ARIMA Forecasting Results:**
   - EDA suggested no significant Auto-regressive or Moving Average forecasting could be done.
   - The model provided future sales predictions with a high error margin.
   - Due to insufficient seasonal data (only 12 months), forecasts were not highly reliable and needed much more data for reliable forecast.

3. Return Analysis:

Return revenue (loss) is **~9.2%** of total sales revenue (896,812.49 out of 9,747,747.93)

Let's analyse it.

1. Products with High Return Counts:

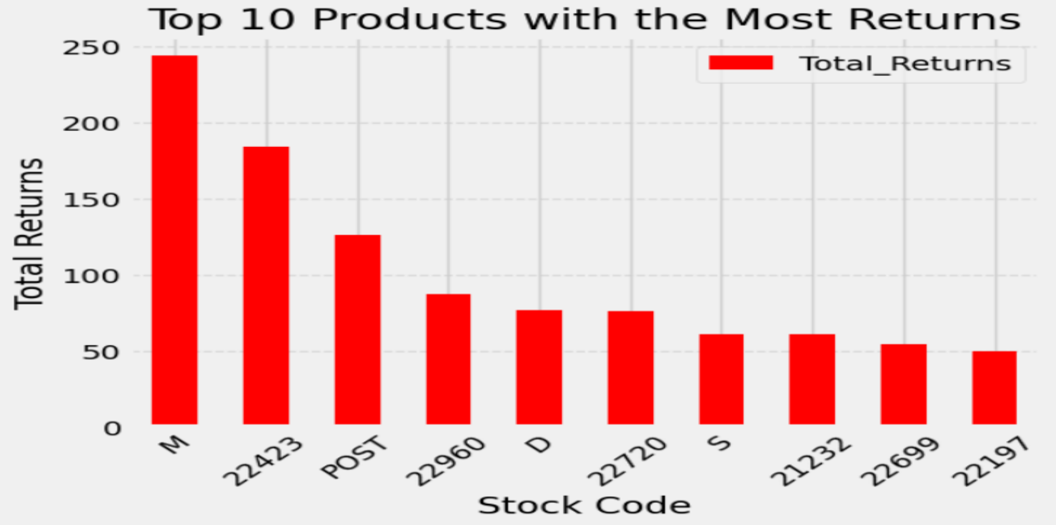


Fig 5.3.1 Top 10 Products with most returns

2. Customers Who Return the Most

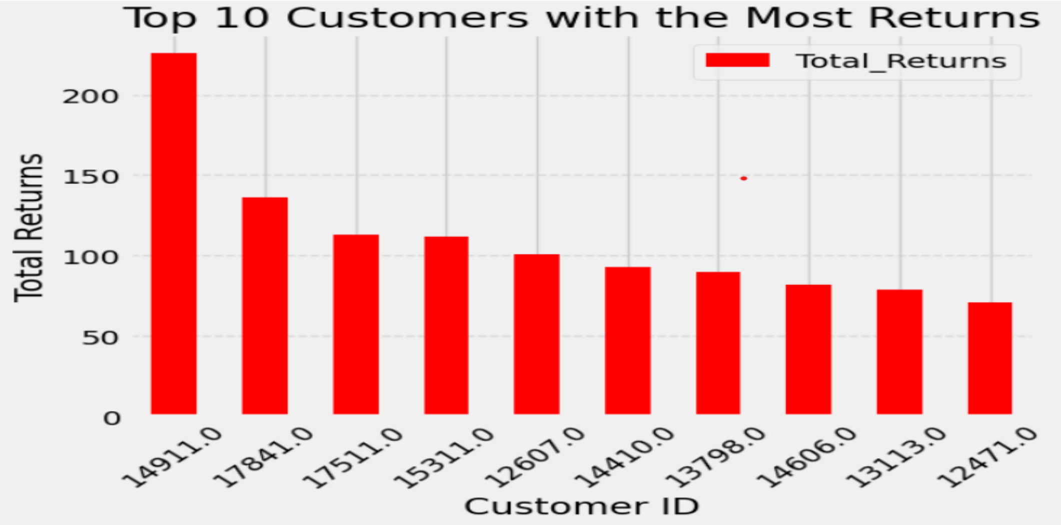


Fig 5.3.2 Top 10 Customers with most returns

3. Top Products Causing Revenue Loss

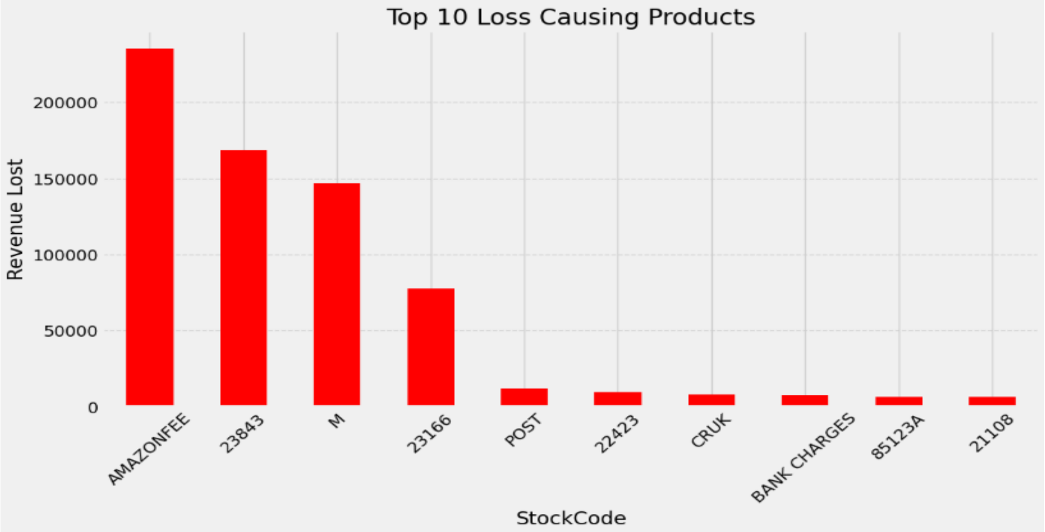


Fig. 5.3.3 Top loss causing products

Observation:

Looks like some values here are the platform charges for amazon, bank etc.

4. Top Customers Causing Revenue Loss (Excluding Unknown_Customer)

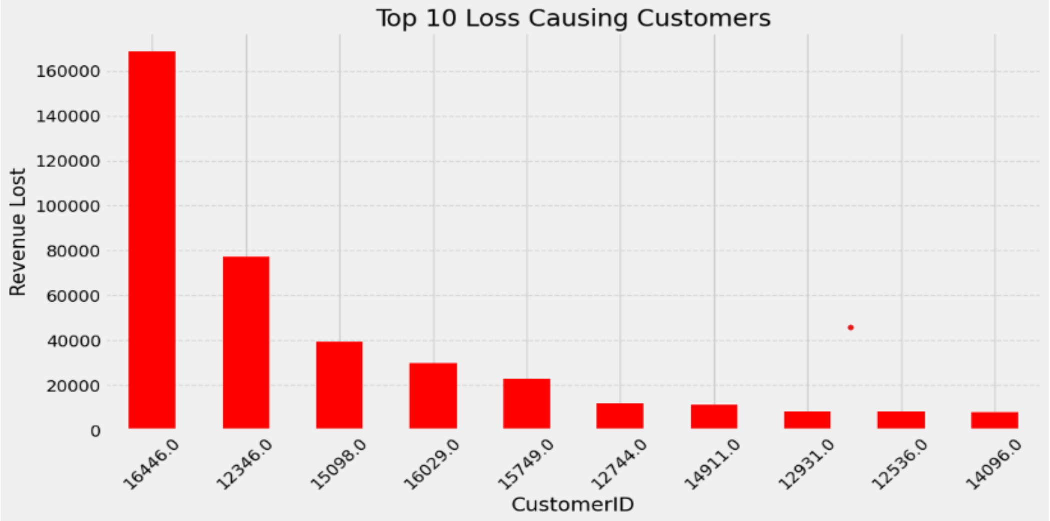


Fig 5.3.4 Top loss causing Customers

5. Return Patterns Over Time

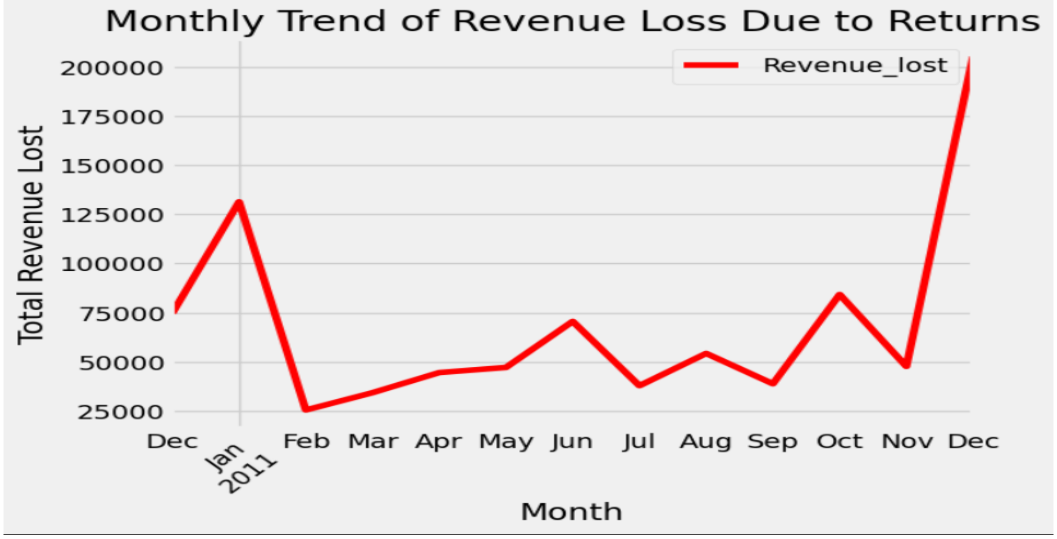


Fig 5.3.5 Return Patterns Over Time (Monthly)

Observation:

Looks like most return losses are generated during November – January period. This could be due to post-holiday returns (common in e-commerce).

6. Country-Wise Return Analysis

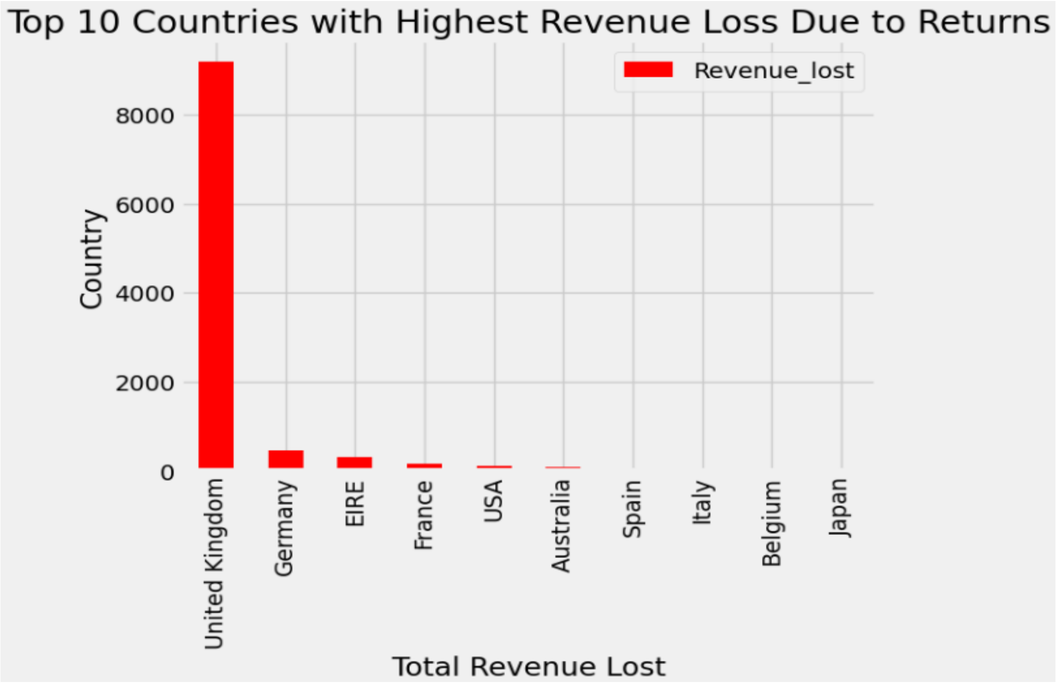


Fig 5.3.6 Top loss generating Country

Observation:

UK generates most losses, but it is the primary market, so this is expected.

4. <u>Customer Segmentation</u>:

After aggregating the Customer with key metrics like **Total Revenue, Total Transaction, Total Products Purchased, Unique Products Purchased** and **Return Frequency** and applying **MinMaxScaling**, the data was ready for clustering.

1. Elbow Method and Silhouette Score to find the best number of clusters for K-Means

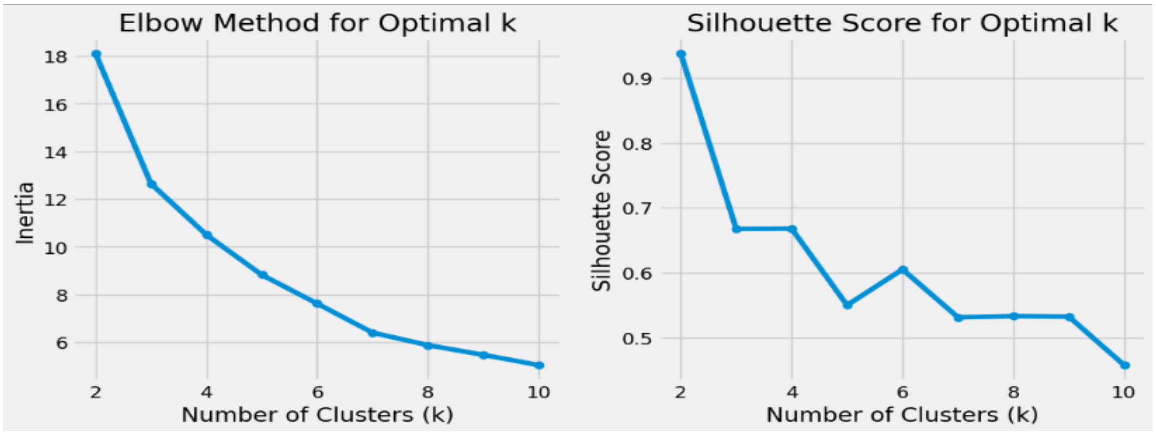


Fig 5.4.1 Elbow method and Silhouette Score for optimal K

Interpretation:

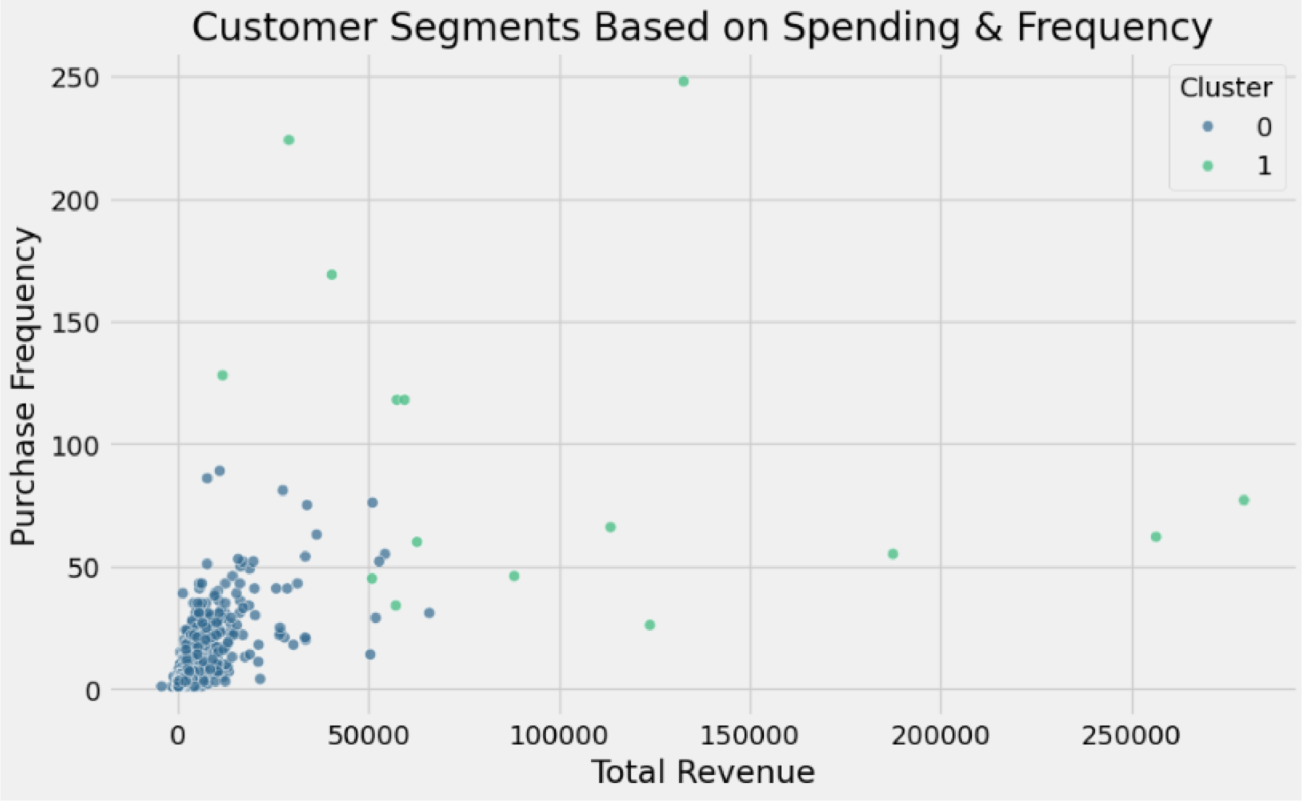
Both Elbow method and Silhouette score suggests that k = 2 is the best number of clusters for K-Means.

2. **After applying K-Means with k=2**

Clusters	**Avg. Total Revenue (GBP)**	**Avg. Total Transactions**	**Avg. Total Products Purchased**	**Avg. Unique Products Purchased**	**Avg. Return Frequency**
0	1549.29	4.75	928.35	58.68	1.84
1	103319.04	98.40	57469.87	794.33	59.60

Table 5.4.1 Cluster Averages

3. **Clusters Visualised**



`

Fig 5.4.2 Customer Segments

6. Interpretation of results and Recommendations

- **Pareto Analysis:**
  1. Top selling products included DOTCOM POSTAGE, Regency Cakestand 3 tier, Paper craft + Little birdie along with 820+ out of 4050+ products. Generated over 20% revenue.

Recommendation: Prioritize high-revenue products for inventory management, marketing, and stock optimization.

2. Top customers included CustomerID 14646, 18102, 17450 along with 930+ out of 4300+ customers (including Unknown Customers). Contributed over 20% revenue.

Recommendations:

1. Implement a loyalty program for high-value customers to maximize retention.
2. Address the "Unknown_Customer" issue by improving data collection on customer identities.
3. Use personalized marketing strategies to increase customer engagement.

- **Demand Forecasting:**

1. **Daily Revenue Trend:** High variability, extreme spikes, and dips indicate irregular demand.

2. **Monthly Revenue Trend:** Cyclical patterns show revenue drops early in the year and peaks in November-December, likely due to the holiday season.

Recommendation:

1. Consider seasonal marketing strategies (e.g., holiday discounts) to capitalize on peak months.
2. Investigate low-revenue months for potential operational improvements.

3. **ARIMA Model Performance:**
   - **MAE**: 320,301.66
   - **MSE**: 128,150,332,679.47
   - **RMSE**: 357,980.91
   - **Interpretation**:
     1. The model's error margin is too high, confirming that time-series forecasting is not suitable for this dataset.
     2. Not enough information is provided. For effective demand forecasting, at least 2-3 years of data may yield better results.

Recommendation:

- Forecasting should be reattempted with external factors (seasonality, holidays, promotional events) and more data (at least 2-3 years).
- Machine learning-based regression models (Random Forest, XGBoost) might perform better.

- **Return Analysis:**

Key Findings

- **Total Revenue Lost Due to Returns:** GBP **896,812.49** (~9.2% of total sales).
- **Peak Return Period: November – January**, likely post-holiday returns.
- **Top Loss-Generating Customers:** A small subset of customers drives high returns.
- **UK accounts for the highest return losses**, but this is expected as it is the primary market

Recommendation:

1. **Improve product quality control** for high-return items.
2. Implement **stricter return policies** during the holiday season.
3. Offer **discounts on non-returnable items** to customers with high return rates.
4. **Analyze return reasons** to reduce avoidable returns (e.g., wrong product descriptions).

- **Customer Segmentation:**
  K-Means (K=2) was the best clustering approach based on the Elbow Method & Silhouette Score.

  Cluster 0 – Regular Shoppers

  - Majority of customers fall in this group.
  - They make occasional purchases with lower total spending.
  - Fewer product returns (Return Frequency = 1.84).

  Recommendations:

  - Can be targeted with promotions & loyalty programs to increase engagement.

  Cluster 1 – High-Value Customers

  - These customers spend significantly more.
  - They make frequent purchases and buy a wide variety of products.
  - Higher return frequency (59.6), indicating potential dissatisfaction or bulk purchasing behaviour.

  Recommendations:

  - Can be offered premium services (exclusive deals, faster shipping, or VIP support).
  - High return rates may require stricter return policies or improved quality assurance.

7. Overall Recommendations Summary

- **Better Data Management:** Improve data collection on customer identities. Asking Reasons for returns.
- **Inventory Optimization:** Focus on the top 20% of revenue-generating products to improve stock efficiency.
- **Customer Retention:** Use loyalty programs and personalized marketing for high-value customers.
- **More Data:** Reattempt demand forecasting with at least 2-3 years of time series data.
- **Forecasting Alternatives:** Consider ML-based models instead of ARIMA due to poor autocorrelation.
- **Return Reduction:** Investigate high-return products/customers and implement stricter policies.
- **Targeted Marketing:** Use segmentation insights to offer promotions for regular shoppers and exclusive perks for high-value customers.