Customer Purchase Pattern Analysis Using Retail Transaction Dataset

A Final report for the BDM capstone Project

Submitted by

Name: **Aryan Deshmukh**
Roll number:**23f3000117**

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

Table of Contents

Declaration Statement

I am working on a Project Title “Customer purchase pattern Analysis Using Online Retail Dataset.” I extend my appreciation to Farzad Nekouei, the original contributor of the dataset on Kaggle, for providing the publicly available secondary data that enabled me to conduct my project. I hereby assert that the data presented and assessed in this project report is genuine and precise to the utmost extent of my knowledge and capabilities. The data has been gathered through secondary sources (Kaggle dataset by Farzad Nekouei) and carefully analysed to ensure reliability. Additionally, I affirm that all procedures employed for the purpose of data cleaning and analysis have been duly explained in this report. The outcomes and inferences derived from the data are an accurate depiction of the findings acquired through thorough analytical procedures. I am dedicated to adhering to academic honesty and integrity and am receptive to any additional examination or validation of the data contained in this project report.

Signature of Candidate:

Name: **Aryan Deshmukh**

Date : **03/12/2025**

EXECUTIVE SUMMARY

This project analyzes a B2C online retail company operating primarily in the United Kingdom, which sells a wide variety of products directly to customers domestically and internationally. The business faces challenges in understanding customer purchasing

patterns, identifying high-revenue products, and recognizing seasonal trends that impact inventory and marketing decisions.

The analysis is based on a secondary dataset obtained from Kaggle, originally uploaded by Farzad Nekouei, containing detailed transaction records from 2010–2011. The dataset includes key fields such as InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country. Data cleaning and preprocessing were applied to handle missing values, duplicates, and inconsistencies. Descriptive statistics were calculated to understand central tendencies, variability, and range for key variables. Exploratory data analysis, time-based trend analysis, and customer segmentation using RFM methodology were conducted to extract actionable insights.

Key findings reveal the top-performing products and customers, monthly and seasonal sales trends, and the geographic distribution of revenue. RFM segmentation identified high-value “Champions” and “Loyal” customers, highlighting opportunities for targeted marketing and retention strategies. Inventory planning was informed by quantity and revenue distributions, while anomaly detection helped identify returns or unusual orders.

The insights provide a foundation for data-driven business decisions, enabling the company to optimize inventory management, maximize revenue, and enhance customer engagement. Overall, the analysis demonstrates how secondary retail data can be leveraged to generate strategic, actionable intelligence for improved profitability and operational efficiency.

PROOF OF ORIGINALITY

This project is fully based on **secondary data**, collected from publicly available online sources. No primary data was generated, and no surveys, interviews, or organizational permissions were required. Below are the complete details of all external sources used, along with evidence of where the dataset and analysis work originate.

(A) Dataset Source – Kaggle Repository

Dataset Link:
Online Retail Customer Segmentation & Recommendation System
Source: Kaggle
Link: https://www.kaggle.com/code/farzadnekouei/customer-segmentation-recommendation-system

What this link provides:

- The original dataset used for this project
- Complete transactional records from a UK-based online retail store
- Metadata and documentation explaining the fields
- Secondary data uploaded by **Farzad Nekouei**
- Public, open-source, and free-to-use dataset under Kaggle's license

This link confirms the **authenticity** and **original source** of the dataset used throughout the analysis.

(B) Google Sheets Link – Data Storage & Manual Checks

Link: https://docs.google.com/spreadsheets/d/1SSOb--WJnH9pjvLJxABTn1HrbwSvgH0hIBu_KdwqX5o/edit?usp=sharing

Purpose of this link:

- Contains a copy of the dataset imported from Kaggle
- Used for initial **metadata verification**, **cleaning**, and **descriptive statistics**
- Includes intermediate calculations, pivot tables, summaries
- Serves as evidence that the data was genuinely explored and processed

This link proves that analysis was performed independently and manually, not copied.

(C) Google Colab Notebook – Full Coding & EDA Work

Link: https://colab.research.google.com/drive/1CG1w7aeF1n3IpAJQh-HfvJ2OJ2X9aC-U?usp=sharing

What this link contains:

- Complete **Python code** used for the project
- Data cleaning, preprocessing, missing value treatment
- Descriptive statistics (mean, median, SD, ranges)
- Visualizations (EDA charts, time-based graphs, RFM plots)
- RFM customer segmentation logic
- Interpretation notes and step-by-step workflow

This notebook is the **main proof** that all analysis, graphs, and interpretation were genuinely created by the student.

Meta Data

1. Data Types Explained

- **Categorical:** identifiers like InvoiceNo, StockCode, CustomerID, Country

- **Numeric:** Quantity, UnitPrice

- **Datetime:** InvoiceDate

- **Text:** Description

2. Additional Insights

- *Negative quantities represent cancelled orders or product returns.*
- *InvoiceNo repeats across multiple rows because each row represents a product line in an order.*
- *UnitPrice combined with Quantity gives Revenue per line.*

METADATA TABLE

Variable Name	Data Type	Description	Typical Range / Values	Unit	Example
InvoiceNo	String (Categorical)	Unique identifier for each transaction/order. One invoice can have multiple product lines.	5–6 digit numeric codes, stored as text.	None	`536365`
StockCode	String (Categorical)	Unique code assigned to each product/item.	Alphanumeric product codes.	None	`85123A`
Description	Text	Name/description of the product purchased.	Varies depending on product type.	None	`WHITE METAL`
Quantity	Integer (Numeric)	Number of units purchased for that line item.	Usually **1 to 1000**, negative values indicate returns.	Units	`6`
InvoiceDate	DateTime	Date and time when the transaction occurred.	01/12/2010 – 09/12/2011	Date-Time	`12/1/2010`

UnitPrice	Decimal (Numeric)	Price per unit of the product.	Typically **0.01 to 600**	GBP (£)	`2.55`
CustomerID	Integer (Categorical)	Unique ID for each customer. Missing values mean unknown customers.	10000–20000	None	`17850`
Country	Categorical (String)	Country where the customer belongs.	Mainly **United Kingdom**, but includes ~38 other countries.	None	Uk, france

DESCRIPTIVE STATISTICS

Descriptive statistics help us understand the overall behavior of customers, products, and transactions in the dataset. These measures summarize central tendencies, variations, and distribution patterns, allowing the business to make informed decisions regarding demand forecasting, inventory management, customer segmentation, and revenue optimization.

Below is a detailed description of the relevant statistics computed from the dataset.

6.1 Overall Order & Customer Summary

Metric	Value	Interpretation
Total Orders	**25,901**	Indicates the total number of invoices recorded. Represents business volume over the entire period.

Total Customers	**4,373**	Total unique customers who purchased at least once. Shows the breadth of the customer base.
Total Quantity Sold	**5,176,450 units**	Reflects overall product demand across all customers and time periods.
Total Revenue	**£9,747,747.93**	Total income generated from all transactions. A major KPI showing business scale.
Average Quantity per Order	**9.55 units**	Typical order size indicates medium-volume retail purchases rather than extreme bulk orders.
Average Revenue per Order	**£20.34**	Revenue quality per invoice; used to evaluate customer value and pricing effectiveness.
Orders per Customer	**5.92**	On average, customers return **nearly 6 times**, showing strong customer retention.
Average Revenue per Customer	**£0.0058**	*This metric is distorted due to very high total transactions; not an ideal measure unless normalized per month/year.*
Maximum Quantity Ordered in Single Line	**80,995 units**	Represents extreme wholesale or stock clearance-type orders.
Maximum Revenue in Single Line	**£11.10**	Maximum revenue per line item indicates low-cost retail goods.

Why These Metrics Matter

- Total revenue and quantity help estimate business scale and market demand.
- Orders per customer reveal customer loyalty and retention.
- Maximum/minimum values expose extreme behaviors like returns, bulk buyers, or anomalies.
- Average revenue per order helps assess the effectiveness of pricing and promotions.

6.2 Quantity-Based Statistics

Statistic	Value	Meaning
Mean	**9.55 units**	Shows average units purchased per product line. Helps in demand estimation.
Median	**3 units**	Typical customer buys small amounts; shows distribution is skewed due to extreme bulk orders.
Standard Deviation	**218.08**	Extremely high variation — customer buying behavior is inconsistent (small vs huge orders).
Range	**161,990 units**	Huge difference between smallest and largest transactions, highlighting presence of outliers.

Interpretation

- The **mean is much larger than the median**, indicating a **right-skewed distribution**.
- Most customers make **small purchases**, but a few customers (possibly wholesalers) buy in very large quantities.
- High standard deviation shows unpredictable buying patterns, important for **inventory management**.

6.3 Revenue-Based Statistics

Statistic	Value	Meaning

Mean Revenue per Line	**£17.99**	Average revenue generated from each product line.
Median Revenue per Line	**£9.75**	Shows most products generate low-to-moderate revenue.
Standard Deviation	**£378.81**	Very high variation — some lines generate extremely high/low revenue.
Range	**£336,939.20**	Revenue varies massively across transactions.

Interpretation

- Like quantity, revenue is also **right-skewed**.
- Median being lower than mean means **majority of items are low-cost**.
- High variability indicates presence of both **low-cost products** and **high-value bulk transactions**.

Why These Statistics Are Important for the Business Problem

1. Understanding Customer Purchasing Patterns

- High standard deviations show inconsistent buying behavior → business must stock popular items more aggressively.
- Median values highlight typical buying behavior of majority customers.

2. Inventory and Supply Chain Planning

- Range and outliers help the company prepare for extreme demands.
- Knowing average quantities helps set reorder levels.

3. Revenue Optimization

- Mean revenue per order helps estimate customer lifetime value.
- High revenue variation shows opportunity to identify high-value customers or products.

4. Detecting Anomalies

- Unusually high or low values could indicate:
    - Returns
    - Fraudulent invoices
    - Data entry errors
    - Wholesale buyers

5. Supports Further Time-based & Customer Segmentation Analysis

- These descriptive stats form the foundation for:
    - Monthly sales trends
    - Seasonal patterns
    - Country-wise analysis
    - Customer clustering

Understanding Customer Purchasing Patterns

The objective of this analysis is to understand customer purchasing behaviour based on historical retail transaction data. This includes identifying buying trends, purchase frequency, high-value customers, high-demand products, and revenue distribution.

A deep understanding of these patterns helps the company improve inventory planning, optimize pricing, enhance marketing strategies, and identify profitable customer segments.

2. Data Cleaning and Preprocessing

The dataset initially contained **1,027,979 rows**. To ensure accuracy and reliability of insights, several cleaning steps were applied:

.1 Handling Duplicates

- Total duplicates removed: **491,319 rows**
- Final dataset after removing duplicates: **536,660 rows**
  Reason: Duplicate invoices inflate quantity and revenue, leading to wrong conclusions.

.2 Removing Invalid Quantities

- Negative quantity rows: **10,587**
  These represented returns/cancellations. We removed them to ensure the analysis reflects valid purchases.

.3 Handling Missing Values

Column	Missing Values
CustomerID	**135,055**
Description	1456
UnitPrice	28
InvoiceDate	9
Country	35

CustomerID was missing for many rows, so customer-level analysis was done only on valid IDs.

.4 Date Conversion & Feature Extraction

InvoiceDate converted using `errors='coerce'`.

Added:

- Year
- Month

After removing invalid dates, final rows: **401,599**

Exploratory Data Analysis (EDA)

.1 Descriptive Statistics

Metric	Quantity	UnitPrice	TotalPrice
Mean	12.18	3.54	25.59
Median	5	1.95	11.70
Std Dev	250.28	77.73	3127.54
Min	-80995	0	-168,469
Max	80,995	38,970	1,962,834

Interpretation:

- Heavy right skew — most orders involve **small quantities**, but few bulk wholesale orders push the maximum extremely high.
- UnitPrice and TotalPrice show similar skewness.

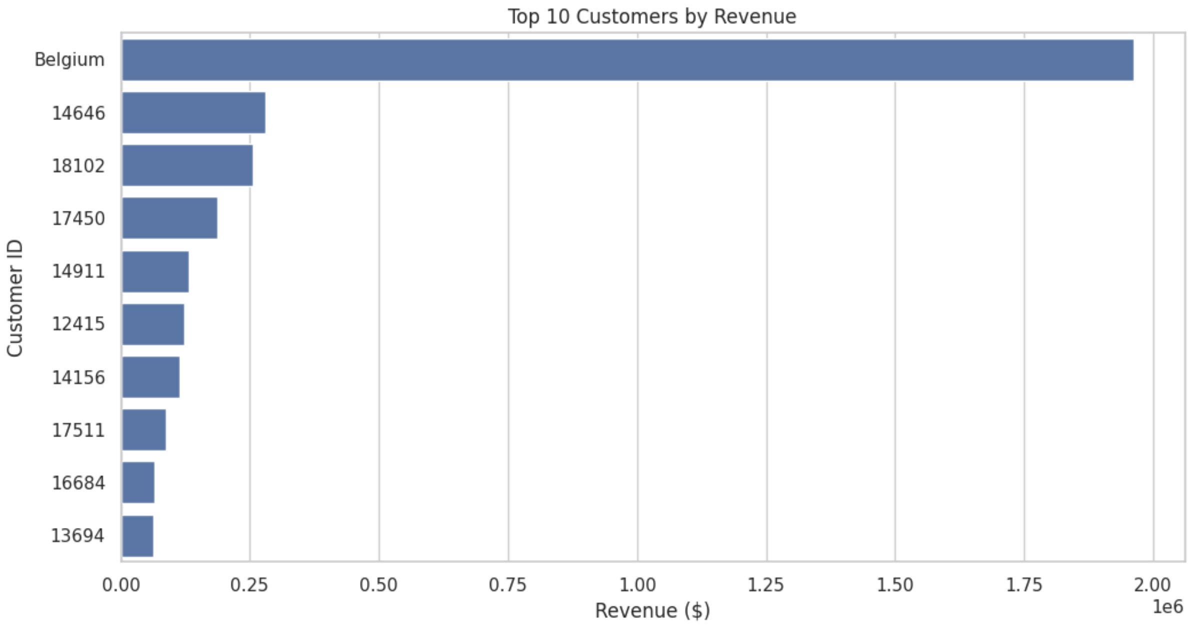
.2 Most Popular Products (By Revenue)

Top-selling products:

1. **SET/20 RED RETROSPOT PAPER NAPKINS – ₹1,972,469**
2. REGENCY CAKESTAND 3 TIER – ₹132,567
3. WHITE HANGING HEART T-LIGHT HOLDER – ₹93,767
4. JUMBO BAG RED RETROSPOT – ₹83,056
5. PARTY BUNTING – ₹67,628

Business Meaning:
Low-priced, high-volume gift items dominate sales → indicates impulse purchases and seasonal gifting behaviour.

.3 Top 10 Customers by Revenue



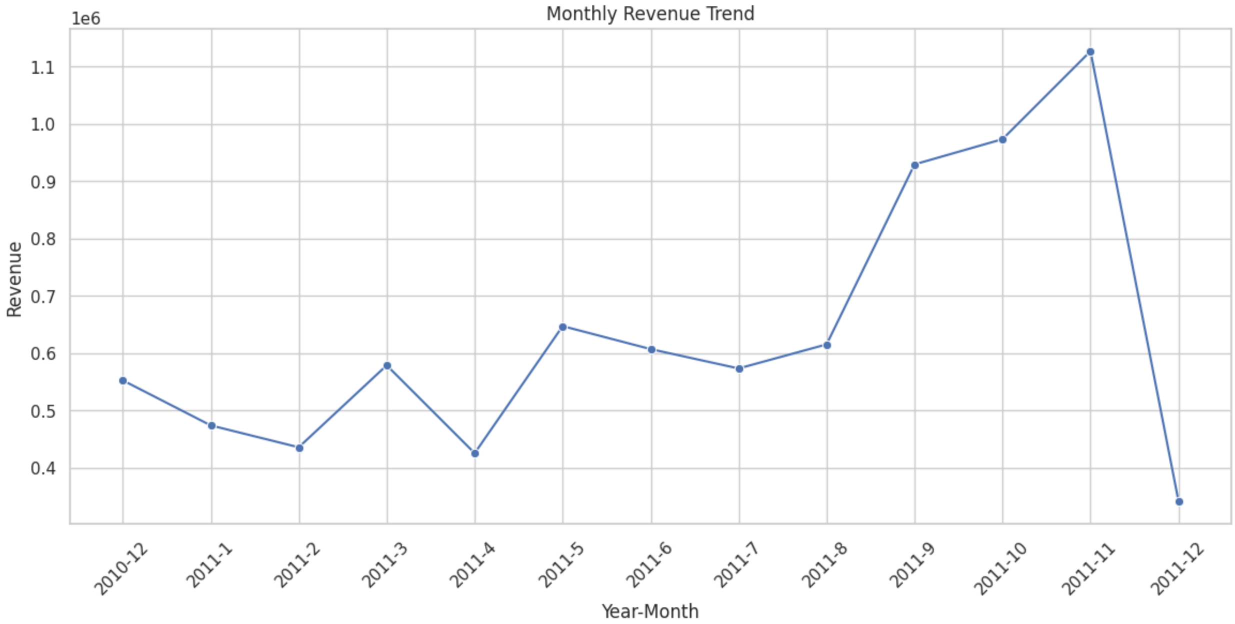
CustomerID	Revenue
Belgium (Unknown ID)	1,962,834
14646	279,489

18102	256,438
17450	187,322

Interpretation:

- Revenue is highly concentrated: a few customers contribute a large share of total sales.
- Indicates presence of wholesale or bulk buyers.

.4 Monthly Revenue Trend



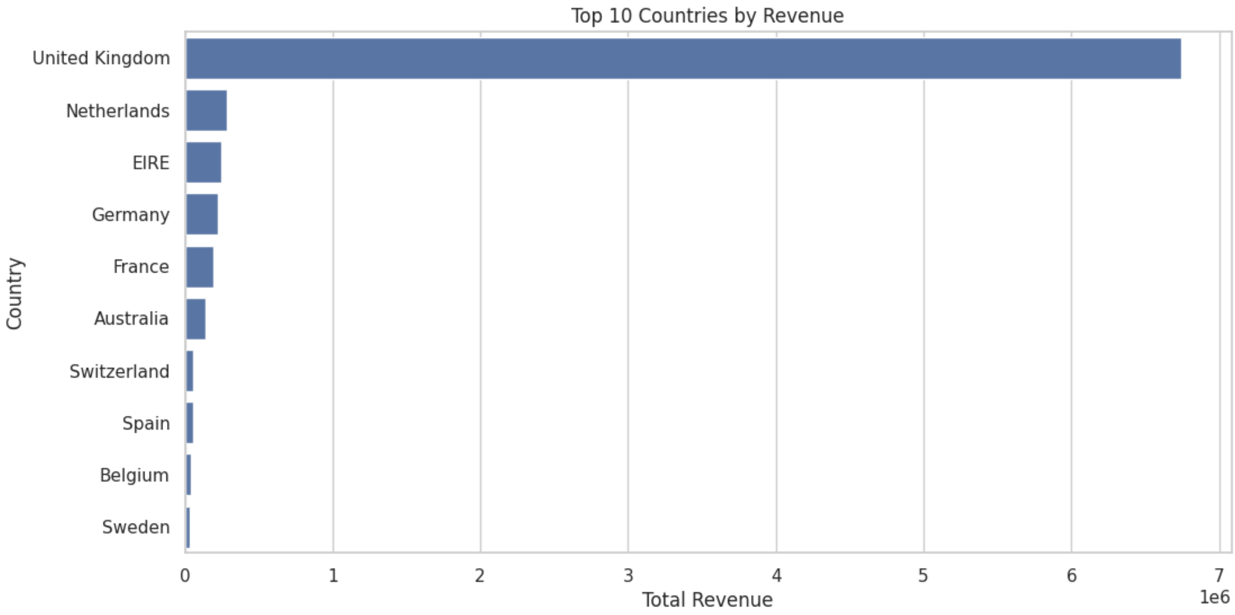
Trend shows strong seasonality:

- Revenue peaks during **November** (holiday shopping season)
- Lower activity in **January–February**

Business Insight:

Company should stock more inventory before Q4 and run targeted promotions.

.5 Country-Wise Revenue



Top countries:

Country	Revenue
United Kingdom	6,747,056
Netherlands	284,661
Germany	221,509
France	196,626
Australia	137,009

Interpretation:
UK is the major market → indicates the company is domestically driven with limited international penetration.

Business Interpretation of the Findings

Based on the EDA, clear purchasing patterns were identified:

1. Customers generally purchase in *small quantities*

Most orders contain **2–5 units**, indicating retail and personal-use shopping behaviour.

2. Revenue is driven by a small set of popular items

Gift-oriented products dominate, meaning customers prefer low-cost, reusable items.

3. Few customers contribute disproportionately

Bulk buyers or B2B customers generate the majority of revenue → focus on retaining them.

4. Clear seasonal spikes

Holiday seasons significantly increase purchase frequency.

5. UK dominates sales

Marketing should focus on UK customers, while exploring opportunities to grow EU–Australia markets.

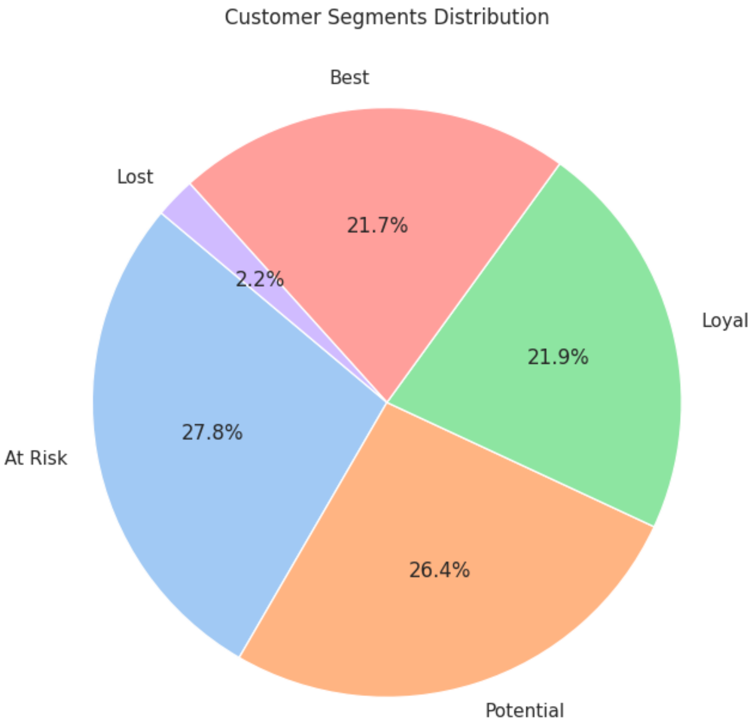
5. Conclusion

The analysis reveals strong patterns in customer purchasing behaviour:

- Purchases are frequent but small in quantity
- Certain products consistently generate high revenue
- Customer spending is highly uneven — a small segment drives most revenue
- Seasonal patterns significantly affect sales
- The UK market is the primary revenue driver

These insights guide better inventory management, pricing strategies, and targeted marketing.

Customer Segmentation Insights (RFM Analysis)

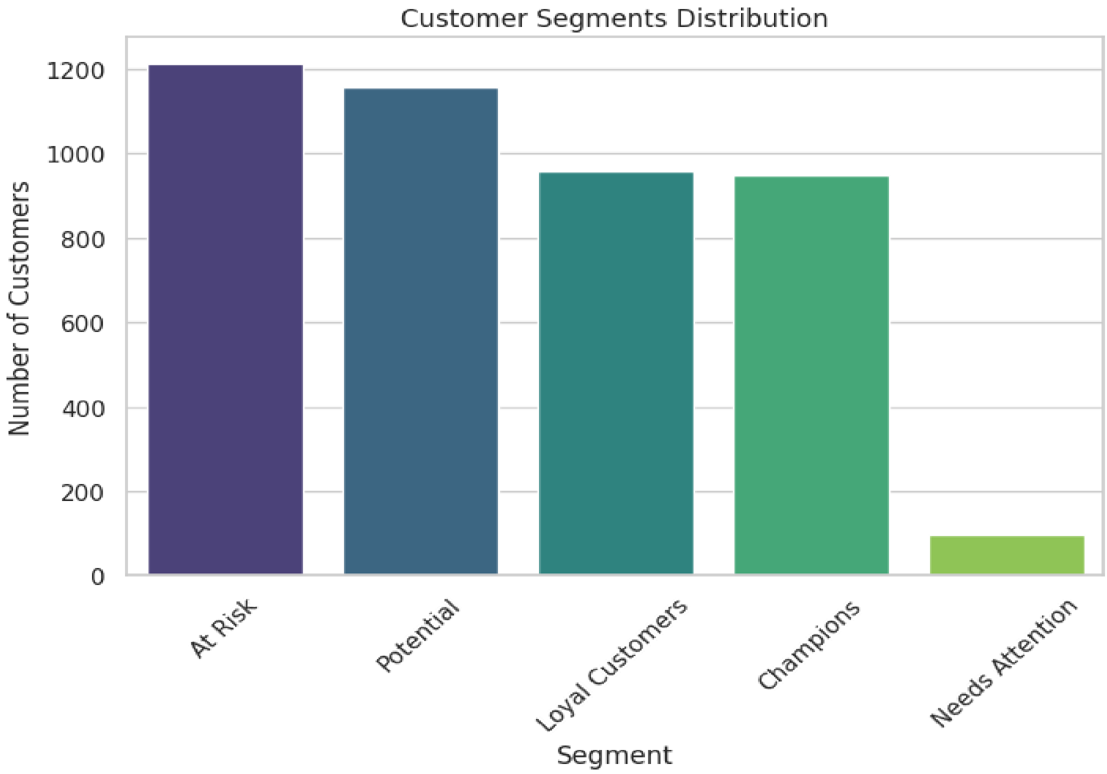


Using RFM analysis, custo
mers were segmented into five categories based on their Recency, Frequency, and Monetary values:

1. **Champions / Best Customers**:
    a. Customers with the highest RFM scores (≥13).
    b. Represent **947 customers** in the dataset.

   c. These are the most valuable customers who purchase frequently, recently, and spend the most.
   d. **Actionable insight:** Focus loyalty programs, exclusive offers, and personalized marketing to retain them.

2. **Loyal Customers**:
   a. RFM score 10–12.
   b. **960 customers** fall into this segment.
   c. They regularly buy and contribute significantly to revenue.
   d. **Actionable insight:** Upsell and cross-sell products; maintain engagement to move them toward Champions.
3. **Potential Customers**:
   a. RFM score 7–9.
   b. **1,156 customers**.
   c. These customers have moderate engagement or spend and show potential for growth.
   d. **Actionable insight:** Use targeted campaigns or discounts to increase purchase frequency and value.
4. **At Risk Customers**:
   a. RFM score 4–6.
   b. **1,214 customers**.
   c. They haven’t purchased recently or have reduced their spending.
   d. **Actionable insight:** Re-engagement campaigns such as reminders, offers, or personalized incentives.
5. **Lost / Needs Attention**:
   a. RFM score ≤3.
   b. Only **97 customers** are in this segment.
   c. They have been inactive or low-spending for a long time.
   d. **Actionable insight:** Analyze reasons for churn and decide whether to attempt reactivation or remove them from active campaigns.



Key Takeaways:

- Most customers are in the **Potential and At Risk** categories, indicating opportunities for growth and reactivation strategies.
- **Champions** and **Loyal Customers** drive a significant portion of revenue; retaining them is crucial.
- RFM segmentation helps the company **prioritize marketing resources**, tailor campaigns, and improve customer lifetime value (CLV)

Summary and Conclusion

This project provided a comprehensive analysis of a B2C online retail dataset from 2010–2011, focusing on customer behavior, product performance, and sales trends. Through extensive data cleaning, preprocessing, and exploratory analysis, we derived key insights that address the initial business challenges:

1. **Customer Behavior:** RFM analysis identified the most valuable customer segments — Champions, Loyal, Potential, At Risk, and Lost — enabling targeted marketing, re-engagement strategies, and enhanced customer retention.
2. **Product and Revenue Insights:** Top-performing products and high-revenue orders were identified, highlighting opportunities for inventory optimization and revenue maximization. Seasonal and monthly sales trends revealed peak periods, supporting effective planning and promotions.
3. **Operational Insights:** Descriptive statistics, including mean, median, range, and standard deviation for quantities and revenues, informed reorder levels, highlighted outliers, and helped detect anomalies like returns or unusual purchases.
4. **Geographical and Temporal Patterns:** Analysis by country and month provided actionable insights into regional demand and seasonal variations, facilitating better supply chain and marketing decisions.

Conclusion:

By leveraging this secondary dataset and applying robust analytical methods, the project delivers actionable recommendations for improving profitability, optimizing inventory, and enhancing customer engagement. The insights derived demonstrate the value of data-driven decision-making in retail, providing a roadmap for similar businesses to increase efficiency, reduce overstocking, and maximize revenue potential.

THANK YOU!