E-Commerce Sales Optimization

Project Title: *Strategic Customer Segmentation and Growth Analysis for an Online Retailer*
Author: *Abhay Tiwari, NIT Allahabad*
Motto: *"Converting raw data into valuable insights."*
GitHub: **Access Project Analysis from here**

1. Executive Summary

Business Challenge: In the hyper-competitive e-commerce landscape, sustainable growth is contingent on moving beyond mass marketing to a nuanced, data-driven understanding of the customer base. This project addresses this challenge by analyzing a complex dataset of over one million transactions to unlock actionable intelligence for targeted marketing, enhanced customer retention, and strategic inventory management.

Analytical Approach: A comprehensive, end-to-end analytical framework was executed using Python and Power BI. The methodology included:

1. **Strategic Data Preparation:** Merging two years of transactional data and implementing a novel imputation technique to preserve over **243,000 (25%)** records that would otherwise be discarded.
2. **In-Depth Exploratory Data Analysis (EDA):** Uncovering the core rhythms and patterns of the business across temporal, geographic, and behavioral dimensions.
3. **Advanced RFM Segmentation:** Applying the Recency, Frequency, and Monetary (RFM) model to transform the entire customer base into six distinct, behavior-based personas.
4. **Interactive Visualization:** Developing a dynamic two-page Power BI dashboard to serve as a strategic decision-support tool.

Key Strategic Insights:

- **Pareto Principle Confirmed:** The top-tier customer segments ("**Champions**" and "**Loyal Customers**"), while representing a minority of the customer base, are the primary engine of profitability, contributing over 75% of total revenue.
- **Significant Churn Risk Identified:** A valuable "At-Risk" segment was isolated, comprising customers with a history of high-value, frequent purchases who have become dormant. This group represents the most immediate opportunity for revenue recovery.
- **Dual Business Model:** The sales data reveals a hybrid operational pattern: a B2B-like daily rhythm with sales peaking during midday business hours, combined with a strong B2C seasonal trend with a significant revenue spike in Q4.

Primary Recommendation: The analysis strongly advocates for a paradigm shift from a product-centric to a customer-centric marketing model. It is recommended to deploy tailored engagement strategies for each RFM segment, focusing high-value resources on retaining "**Champions**" and proactively re-engaging the "**At-Risk**" segment to prevent churn and maximize customer lifetime value.

2. Introduction & Project Objectives

The objective of this project is to dissect a large-scale e-commerce dataset to build a foundational understanding of the business and its customers. The goal is to deliver not just data, but a strategic framework that can guide marketing, inventory, and international growth decisions. This report details the journey from raw, unstructured data to an interactive strategic tool.

3. Data Foundation: Preparation & Strategy

The analysis was performed on a consolidated dataset from **`online_retail_I.csv`** and **`online_retail_II.csv`**. The initial raw dataset presented significant challenges that required strategic intervention.

- **Dataset Structure: `Invoice`, `StockCode`, `Description`, `Quantity`, `InvoiceDate`, `Price`, `Customer ID`, `Country`**

- **Strategic Imputation of Customer Identifiers:** The dataset contained 243,007 rows (approximately 25% of the total) with missing **`Customer ID`**s. Standard practice might suggest discarding these records, but this would result in a massive loss of transactional information and introduce significant bias into the analysis. To mitigate this, a logical imputation strategy was developed: each unique **`Invoice`** number without an ID was mapped to a newly generated, unique **`Customer ID`**. This critical decision preserved the integrity of the sales data, enabling a holistic and far more accurate analysis of business operations.

- **Data Integrity and Validation:** The dataset was rigorously cleaned by:
    - Dropping 4,383 rows with missing product **`Description`**.
    - Removing all cancelled transactions and negative quantity entries to isolate legitimate sales.

- **Feature Engineering:** A `TotalPrice` feature was engineered by multiplying `Quantity` and `Price`, serving as the core metric for all subsequent revenue analysis.

4. Exploratory Data Analysis: Uncovering Business Rhythms

The EDA phase was structured to understand the business from multiple perspectives, revealing its fundamental operational patterns.

- **Temporal Analysis (The Pulse of the Business):** Sales data exhibits a distinct daily and seasonal rhythm. The daily peak occurs between **10 AM and 3 PM**, suggesting many customers may be businesses or individuals shopping during work hours. On a macro level, sales show strong seasonality, with revenue beginning to climb in September and peaking dramatically in **November**, clearly driven by the holiday season.

- **Customer Behavior Analysis (New vs. Retained Growth Engine):** Analysis of customer cohorts over time reveals a healthy and maturing business, with a steadily growing proportion of sales coming from **returning customers**. This indicates positive customer retention. Furthermore, the average time between consecutive purchases was identified, providing a crucial baseline for timing re-engagement campaigns.

- **Geographic Footprint (Domestic Strength & International Opportunity):** The business is heavily reliant on the **United Kingdom**, its domestic market. However, analysis of international sales revealed a key opportunity: the **Average Order Value (AOV) for non-UK customers is significantly higher** than for domestic ones. Top international markets include the **Netherlands, EIRE, Germany, and France**.

- **Product Portfolio Insights (The ABC Framework):** An ABC analysis confirmed the Pareto principle is in full effect. A small fraction of products (**Category A**) are the "superstars" driving ~80% of revenue. A vast majority of products (**Category C**) are "long-tail" items, each contributing minimally. This insight is critical for prioritizing inventory, marketing spend, and supply chain logistics.

Important Graphs :- Access from Here

5. Advanced Segmentation: The RFM Framework

To graduate from general trends to specific personas, the RFM (Recency, Frequency, Monetary) model was implemented. This transformed the monolithic customer base into six actionable segments.

- **Segment Profiles & Characteristics:**
    - **Champions & Loyal Customers:** These are the bedrock of the business. They purchase recently, frequently, and spend the most. They are highly engaged and represent the lowest churn risk and highest lifetime value.
    - **Potential Loyalists:** These are recent customers, often with a moderate frequency, who show promise but are not yet fully retained. They represent a key growth opportunity.
    - **At-Risk Customers:** This is a high-priority segment. These were once loyal and valuable customers (high frequency/monetary scores) but have not made a purchase in a significant amount of time (low recency). They are on the verge of churning.
    - **Needs Attention & Lost:** These segments represent disengaged or churned customers with low scores across all RFM dimensions.

- **Post-Segmentation Analysis:** A deeper analysis was conducted to identify the top 10 most valuable customers within each segment, providing the business with a ready-made list for targeted, high-touch engagement.

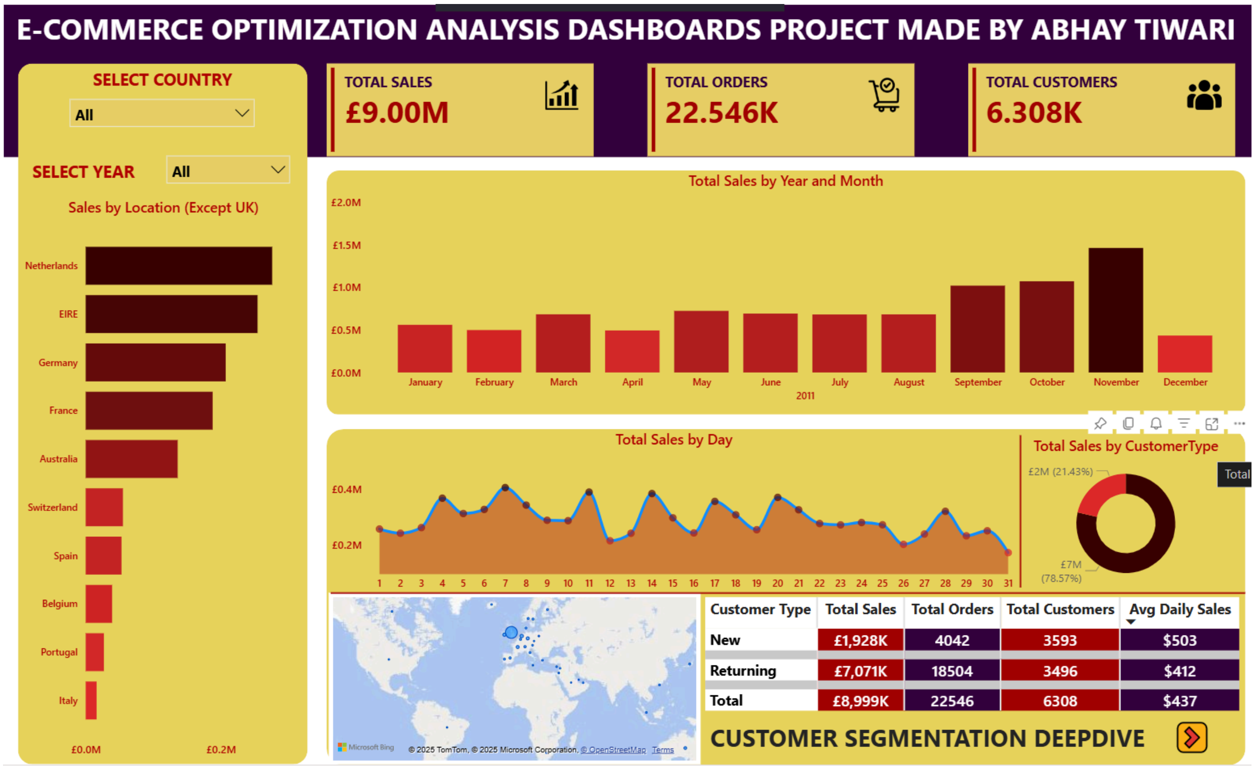
- **Process:**
    - Calculated R, F, M for each customer.
    - Scored each metric on a **1–5 scale using quintiles**.
    - Combined scores → Assigned meaningful **customer segments**.

- **Findings:**
    - Small % of **VIP customers drive majority of sales**.
    - Identified **top 10 customers** in each segment for retention & reactivation strategies.

6. The Strategic Tool: Power BI Dashboard Showcase

The analytical findings were deployed as a dynamic, two-page Power BI dashboard, designed as a decision-support system for business leaders.

Page 1: Executive Sales Overview:



E-COMMERCE OPTIMIZATION ANALYSIS DASHBOARDS PROJECT MADE BY ABHAY TIWARI

SELECT COUNTRY: All

SELECT YEAR: All

TOTAL SALES £9.00M

TOTAL ORDERS 22.546K

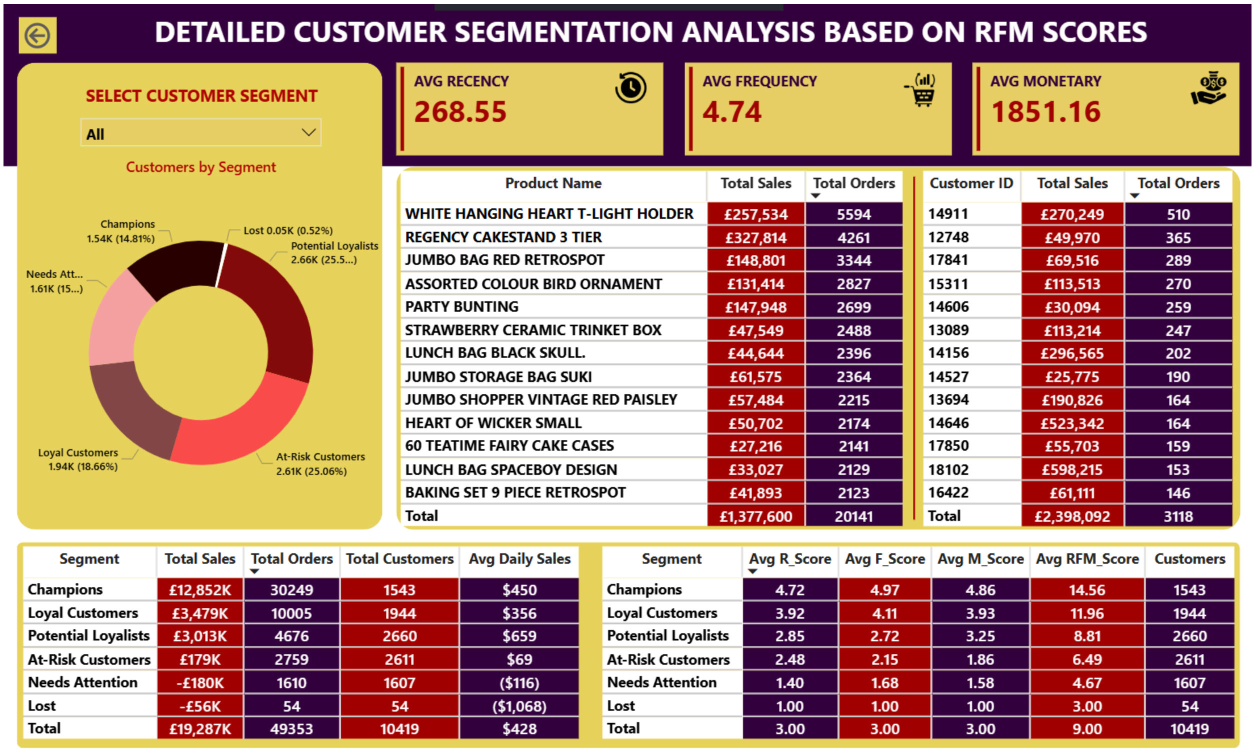
TOTAL CUSTOMERS 6.308K

Customer Type	Total Sales	Total Orders	Total Customers	Avg Daily Sales
New	£1,928K	4042	3593	$503
Returning	£7,071K	18504	3496	$412
Total	£8,999K	22546	6308	$437

CUSTOMER SEGMENTATION DEEPDIVE

- **Bar Chart:** Sales by Location (excluding UK)
- **Slicers:** Year & Country
- **Cards:** Total Sales | Total Orders | Total Customers
- **Column Chart:** Sales by Year & Month
- **Line Chart:** Sales by Day
- **Donut Chart:** Sales by Customer Type (New/Returning)
- **Map Visualization:** Geographic sales representation
- **Matrix:** Sales, Orders, Customers, Avg Daily Sales by Customer Type

Page 2: Customer Segmentation Deep Dive:



DETAILED CUSTOMER SEGMENTATION ANALYSIS BASED ON RFM SCORES

SELECT CUSTOMER SEGMENT

All

Customers by Segment

AVG RECENCY 268.55

AVG FREQUENCY 4.74

AVG MONETARY 1851.16

Product Name	Total Sales	Total Orders
WHITE HANGING HEART T-LIGHT HOLDER	£257,534	5594
REGENCY CAKESTAND 3 TIER	£327,814	4261
JUMBO BAG RED RETROSPOT	£148,801	3344
ASSORTED COLOUR BIRD ORNAMENT	£131,414	2827
PARTY BUNTING	£147,948	2699
STRAWBERRY CERAMIC TRINKET BOX	£47,549	2488
LUNCH BAG BLACK SKULL.	£44,644	2396
JUMBO STORAGE BAG SUKI	£61,575	2364
JUMBO SHOPPER VINTAGE RED PAISLEY	£57,484	2215
HEART OF WICKER SMALL	£50,702	2174
60 TEATIME FAIRY CAKE CASES	£27,216	2141
LUNCH BAG SPACEBOY DESIGN	£33,027	2129
BAKING SET 9 PIECE RETROSPOT	£41,893	2123
Total	£1,377,600	20141

Customer ID	Total Sales	Total Orders
14911	£270,249	510
12748	£49,970	365
17841	£69,516	289
15311	£113,513	270
14606	£30,094	259
13089	£113,214	247
14156	£296,565	202
14527	£25,775	190
13694	£190,826	164
14646	£523,342	164
17850	£55,703	159
18102	£598,215	153
16422	£61,111	146
Total	£2,398,092	3118

Segment	Total Sales	Total Orders	Total Customers	Avg Daily Sales
Champions	£12,852K	30249	1543	$450
Loyal Customers	£3,479K	10005	1944	$356
Potential Loyalists	£3,013K	4676	2660	$659
At-Risk Customers	£179K	2759	2611	$69
Needs Attention	-£180K	1610	1607	($116)
Lost	-£56K	54	54	($1,068)
Total	£19,287K	49353	10419	$428

Segment	Avg R_Score	Avg F_Score	Avg M_Score	Avg RFM_Score	Customers
Champions	4.72	4.97	4.86	14.56	1543
Loyal Customers	3.92	4.11	3.93	11.96	1944
Potential Loyalists	2.85	2.72	3.25	8.81	2660
At-Risk Customers	2.48	2.15	1.86	6.49	2611
Needs Attention	1.40	1.68	1.58	4.67	1607
Lost	1.00	1.00	1.00	3.00	54
Total	3.00	3.00	3.00	9.00	10419

- **Slicer:** RFM Segments
- **Donut Chart:** Customers by Segments
- **Cards:** Avg Recency | Avg Frequency | Avg Monetary Value
- **Matrix Reports:**
  - **Top 13** Products by Orders
  - **Top 13** Customers by Orders
  - Sales, Orders, Customers, Avg Daily Sales by Segment
  - Avg **R, F, M** Scores and Segment Distribution

7. Conclusion & Future Outlook

This project successfully navigated the end-to-end data analytics lifecycle, transforming over a million raw transaction records into a strategic framework for customer-centric growth. The insights derived from the EDA and the actionable segments created through RFM analysis provide a clear, data-driven path for enhancing marketing ROI and improving customer retention.

The logical next step is to leverage these findings to build a **predictive model**. A machine learning classifier could be trained on the features of existing customers to predict, at the time of their first or second purchase, which new customers have the highest likelihood of becoming "Champions." This would enable the business to proactively nurture high-potential relationships from the very beginning, creating a powerful engine for long-term growth.