🛒 Online Retail Customer Segmentation using EDA & RFM Clustering

📌 Project Overview

This project focuses on **Exploratory Data Analysis (EDA)** and **customer segmentation** for an **Online Retail dataset** using **RFM (Recency, Frequency, Monetary) analysis** and **clustering techniques**. The objective is to uncover purchasing patterns, identify valuable customer segments, and generate actionable business insights.

The analysis follows an end-to-end analytics workflow:

1. Data wrangling and cleaning
2. Exploratory Data Analysis (EDA)
3. Feature engineering (RFM metrics)
4. Customer clustering using multiple algorithms
5. Business interpretation of results

📁 Dataset Description

The dataset contains transactional data from an online retail store with the following columns:

Column Name	Description
InvoiceNo	Unique invoice number (each invoice = one transaction)
StockCode	Product identifier
Description	Product description
Quantity	Number of items purchased
InvoiceDate	Date and time of transaction
UnitPrice	Price per unit
CustomerID	Unique customer identifier
Country	Customer's country

🧹 Data Wrangling & Cleaning

To ensure analytical accuracy, the following cleaning steps were performed:

- Converted `InvoiceDate` to datetime format

- Removed records with missing `CustomerID`
- Excluded cancelled transactions (InvoiceNo starting with "C")
- Removed records with negative or zero `Quantity` and `UnitPrice`
- Created a new feature `TotalAmount = Quantity × UnitPrice`

These steps ensure that only **valid, revenue-generating transactions** are used for analysis.

🔍 Exploratory Data Analysis (EDA)

Key EDA Activities

- Distribution analysis of **Quantity**, **UnitPrice**, and **TotalAmount**
- Time-based analysis: sales trends by **month**, **weekday**, and **hour**
- Country-level analysis: revenue and transaction concentration
- Product-level analysis: top-selling and highest-revenue products
- Customer-level analysis: spending concentration and long-tail behavior

Key EDA Insights

- Revenue distribution is **highly skewed**, where a small percentage of customers generate a large portion of total sales
- Sales exhibit **seasonality**, with noticeable monthly and weekday patterns
- The majority of transactions come from a few key countries
- High-value transactions are driven by bulk purchases and premium-priced products

These insights guided the feature selection and scaling strategy for clustering.

📐 Feature Engineering – RFM Metrics

RFM analysis was used to quantify customer behavior.

🔁 Recency

Definition: Number of days since a customer's last purchase

Calculation:

$$\text{Recency} = \text{Latest Invoice Date in Dataset} - \text{Customer's Last Purchase Date}$$

- Lower Recency → more active customers
- Higher Recency → dormant or churn-risk customers

🔄 Frequency

Definition: Number of transactions made by a customer

Calculation:

Frequency = Count of unique InvoiceNo per CustomerID

- One invoice represents one transaction
- Higher Frequency → loyal or repeat customers

💰 Monetary

Definition: Total amount spent by a customer

Calculation:

Monetary = Σ (Quantity × UnitPrice) per CustomerID

- Higher Monetary → high-value / VIP customers
- Monetary values are highly skewed, requiring scaling

📊 RFM Dataset Preparation

After calculating Recency, Frequency, and Monetary:

- Metrics were merged into a single RFM table
- Features were **standardized** using scaling techniques
- The scaled RFM features were used as input for clustering algorithms

---

🤖 Clustering Methodology

Algorithms Used

1️⃣ K-Means Clustering (Benchmark)

- Used to understand overall cluster tendency
- **Elbow Method** applied to identify optimal number of clusters
- **Silhouette Score** used to evaluate cluster separation

2️⃣ Hierarchical Clustering

- Agglomerative clustering with Ward linkage
- **Dendrogram** used to visually determine optimal clusters
- Silhouette score used for validation

3️⃣ DBSCAN

- Density-based clustering (no predefined cluster count)
- Automatically identifies **outliers and noise customers**
- `eps` parameter selected using **k-distance plot**
- Silhouette score calculated after excluding noise points

📈 Model Evaluation

Metric	Purpose
Elbow Method	Identify optimal cluster count (K-Means reference)
Dendrogram	Visual cluster cut-off (Hierarchical)
Silhouette Score	Measure cluster cohesion and separation

🧠 Business Interpretation of Clusters

Based on RFM clustering, customers can be segmented into:

- **Champions**: Low Recency, High Frequency, High Monetary
- **Loyal Customers**: Medium Recency, High Frequency
- **Potential Loyalists**: Recent but lower spending
- **At-Risk Customers**: High Recency, previously active
- **Dormant / Churned**: High Recency, Low Frequency, Low Monetary

DBSCAN further highlights **extreme high-value customers** as outliers, useful for VIP targeting.

🎯 Business Value & Use Cases

- Personalized marketing and promotions
- Customer retention and churn prevention
- VIP customer identification
- Inventory and demand planning
- Strategic decision-making using customer lifetime value

🛠️ Tools & Technologies

- Python (Pandas, NumPy)
- Matplotlib & Seaborn (Visualization)
- Scikit-learn (Clustering & evaluation)
- Jupyter Notebook

☑️ Conclusion

This project demonstrates how **EDA combined with RFM-based clustering** can transform raw transactional data into actionable customer intelligence. By leveraging multiple clustering techniques and robust evaluation metrics, the analysis delivers both **statistical rigor and business relevance**.
