Project Title	🛒 **Shopper Spectrum: Customer Segmentation and Product Recommendations in E-Commerce**
Skills take away From This Project	• Public Dataset Exploration and Preprocessing • Data Cleaning and Feature Engineering • Exploratory Data Analysis(EDA) • Clustering Techniques • Collaborative Filtering-based Product Recommendation • Model Evaluation and Customer Segmentation Interpretation • Streamlit
Domain	E-Commerce and Retail Analytics

📣 Problem Statement

The global e-commerce industry generates vast amounts of transaction data daily, offering valuable insights into customer purchasing behaviors. Analyzing this data is essential for identifying meaningful customer segments and recommending relevant products to enhance customer experience and drive business growth. This project aims to examine transaction data from an online retail business to uncover patterns in customer purchase behavior, **segment customers** based on **Recency, Frequency, and Monetary (RFM) analysis**, and develop a product **recommendation system** using **collaborative filtering** techniques.

📌 Real-time Business Use Cases:

- Customer Segmentation for Targeted Marketing Campaigns
- Personalized Product Recommendations on E-Commerce Platforms
- Identifying At-Risk Customers for Retention Programs
- Dynamic Pricing Strategies Based on Purchase Behavior
- Inventory Management and Stock Optimization Based on Customer Demand Patterns

---

🧠 Problem Type:

- Unsupervised Machine Learning – Clustering
- Collaborative Filtering – Recommendation System

🔧 Project Tasks

Step 1:Dataset Collection and understanding

- Dataset - **Link**
- Explore the dataset to understand the structure and data types.
- Identify missing values, duplicates, and unusual records.

📌Dataset Description

Column	Description
`InvoiceNo`	Transaction number
`StockCode`	Unique product/item code
`Description`	Name of the product
`Quantity`	Number of products purchased
`InvoiceDate`	Date and time of transaction (2022–2023)
`UnitPrice`	Price per product
`CustomerID`	Unique identifier for each customer
`Country`	Country where the customer is based

Step 2: 📌 Data Preprocessing:

- Remove rows with missing CustomerID
- Exclude cancelled invoices (InvoiceNo starting with 'C')

- Remove negative or zero quantities and prices

Step 3 : 📌 Exploratory Data Analysis (EDA):

- Analyze transaction volume by country
- Identify top-selling products
- Visualize purchase trends over time
- Inspect monetary distribution per transaction and customer
- RFM distributions
- Elbow curve for cluster selection
- Customer cluster profiles
- Product recommendation heatmap / similarity matrix

Step 4 : 📌 Clustering Methodology:

1️⃣ Feature Engineering:

- Calculate **Recency** = Latest purchase date in dataset − Customer's last purchase date
- Calculate **Frequency** = Number of transactions per customer
- Calculate **Monetary** = Total amount spent by customer

2️⃣ Standardize/Normalize the RFM values

3️⃣ Choose Clustering Algorithm (KMeans, DBScan, Hierarchial etc)

4️⃣ Use Elbow Method , Silhouette Score to decide the number of clusters

5️⃣ Run Clustering

Label the clusters by interpreting their RFM **averages**:

Cluster	Characteristics	Segment Label
High R, High F, High M	Regular, frequent, recent, and big spenders	**High-Value**
Medium F, Medium M	Steady purchasers but not premium	**Regular**
Low F, Low M, older R	Rare, occasional purchases	**Occasional**
High R, Low F, Low M	Haven’t purchased in a long time	**At-Risk**

6️⃣ Visualize the clusters using a scatter plot or 3D plot of RFM scores.

7️⃣ Save the best performing model for streamlit usage

📌 Recommendation System Approach:

- Use **Item-based Collaborative Filtering**
- Compute **cosine similarity** (or another similarity metric) between products based on purchase history (CustomerID–StockCode matrix)
- Return top 5 similar products to the entered product name

📱 Streamlit App Features

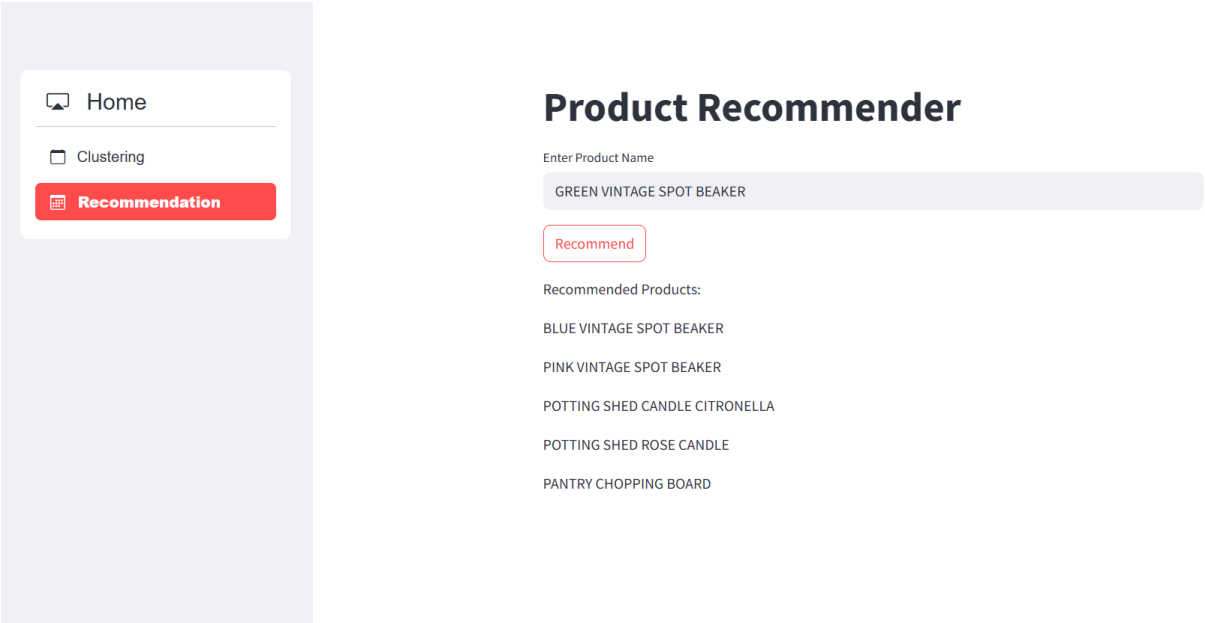
🎯1️⃣Product Recommendation Module

Objective:

When a user inputs a product name, the app recommends **5 similar products** based on collaborative filtering.

Functionality:

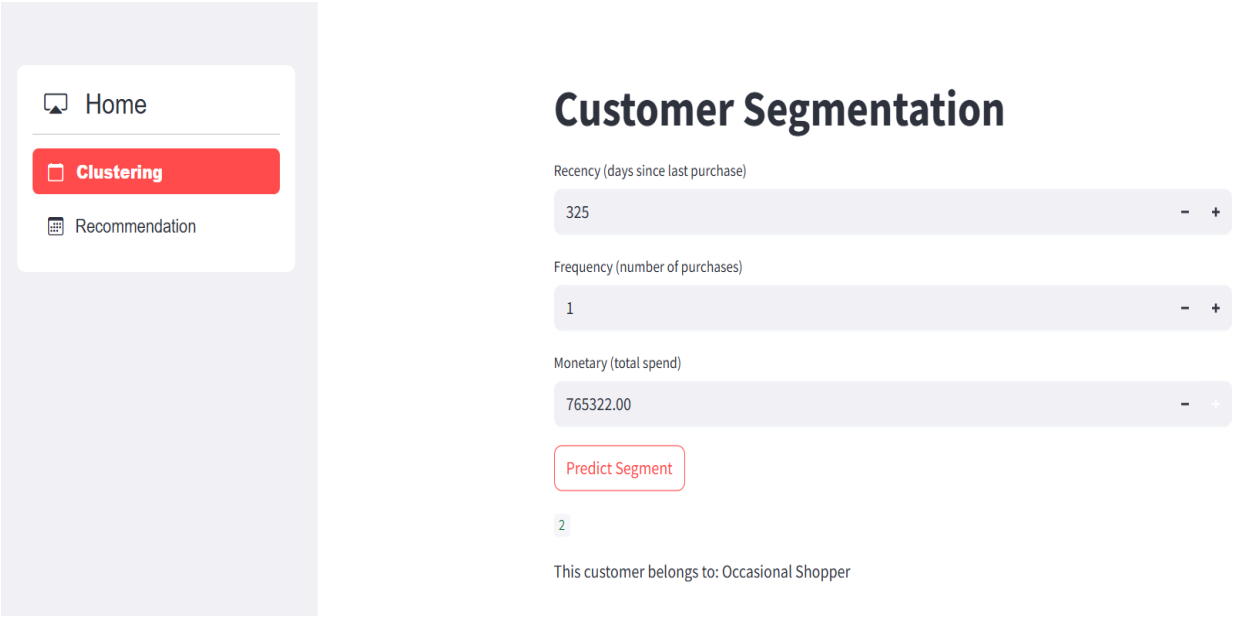
- Text input box for **Product Name**
- Button: `Get Recommendations`
- Display 5 recommended products as a styled list or card view



🎯2️⃣ Customer Segmentation Module

🔍 Functionality:

- 3 number inputs for:
    - **Recency (in days)**
    - **Frequency (number of purchases)**
    - **Monetary (total spend)**
- Button: `Predict Cluster`
- Display: Cluster label (e.g., `High-Value`, `Regular`, `Occasional`, `At-Risk`)



🛠️ Technical Tags

`Pandas,Numpy,DataCleaning,FeatureEngineering, EDA , RFMAnalysis, CustomerSegmentation, KMeansClustering, CollaborativeFiltering, CosineSimilarity, ProductRecommendation, ScikitLearn, StandardScaler, StreamlitApp, MachineLearning, DataVisualization, PivotTables, DataTransformation, RealTimePrediction`

📌 Project Deliverables:

- 📓 Python Notebook with:
    - Clean, well-documented code with comments

    - Visualizations for EDA and clustering insights
    - RFM-based customer segmentation and product similarity analysis
    - Model evaluations for clustering (like inertia, silhouette score)
- 📊 Streamlit Web Application:
    - User input for a product name → recommends 5 similar products
    - Customer behavior input (Recency, Frequency, Monetary) → predicts cluster segment
    - Clean, interactive UI with real-time outputs
- Recorded Video

⏳ Timeline

The project should be completed and submitted **within 7 days** from the date it is assigned.

References

Streamlit recording (English)	Special session for STREAMLIT(11/08/2024)
Streamlit Reference doc	Streamlit API reference
Project Live Evaluation	Project Live Evaluation

Capstone Explanation Guideline	Capstone Explanation Guideline
GitHub Reference	How to Use GitHub.pptx
Machine Learning(Eng) Recommendation systems	Project Excellence Series: Guided Learning & Problem Solving [Machine Learning USVL](English)
Machine Learning(Tam) Recommendation systems	Project Excellence Series: Guided Learning & Problem Solving [Machine Learning USVL](TAMIL)
Project Orientation	xit-tvjx-xth (2025-06-16 11:53 GMT+5:30)