Aura AI - A Retail Analytics Dashboard

Name: Arya Jain

Batch: 01 June Batch

Duration: 6 Months

Course: Data Science/AIML

ABSTRACT

This project, titled "Aura AI," presents an interactive retail analytics dashboard developed using the Streamlit framework. The primary goal is to provide a practical and accessible application of data science and machine learning to key business challenges in the retail sector. The dashboard focuses on two core functionalities: customer segmentation and product recommendation.

Utilizing a dataset of transactional retail data, the application first employs a Recency, Frequency, Monetary (RFM) analysis to quantify customer behavior. A K-Means clustering algorithm is then applied to these RFM metrics to segment the customer base into distinct, behaviorally-defined groups. The dashboard allows users to input new RFM values to predict which of these segments a customer belongs to, providing valuable insights into their purchasing patterns.

For product recommendations, the system implements an item-based collaborative filtering model. By calculating product-to-product similarities based on purchase history, the application can suggest similar products when a user inputs a specific item. This feature aids in identifying cross-selling opportunities and enhancing the customer experience.

The entire application is encapsulated in a user-friendly web interface with a custom aesthetic design. This project serves as a robust demonstration of how machine learning can be leveraged to transform raw sales data into strategic, actionable business intelligence, ultimately enabling better decision-making for customer engagement, marketing, and inventory management.

INTRODUCTION

BACKGROUND

The retail industry operates in a highly competitive and data-rich environment. Businesses are constantly seeking innovative ways to understand customer behavior, predict purchasing trends, and tailor their strategies to maximize profitability. Traditional one-size-fits-all marketing approaches are often inefficient. A more effective strategy involves leveraging data to identify and cater to specific customer segments. Furthermore, enhancing the customer experience through personalized product recommendations is a proven method for increasing sales and customer loyalty. This project addresses these industry challenges by developing a comprehensive analytics platform that utilizes modern data science and machine learning techniques to provide a clear, data-driven perspective on a company's customer base and product catalog.

PROJECT OBJECTIVES

The successful completion of this project resulted in the development of "Aura AI," a functional and user-friendly web application with the following key deliverables:

- **A Streamlit Web Application:** A fully deployed and interactive dashboard that serves as the main user interface for all analytical features.
- **Customer Segmentation Module:** A robust feature that uses a pre-trained K-Means clustering model on RFM (Recency, Frequency, Monetary) metrics to segment customers. This module can predict the cluster of a new customer based on their inputs and provides a detailed profile for each segment.
- **Product Recommendation Engine:** A module that utilizes an item-based collaborative filtering model to generate product recommendations. Users can input a product name to receive a list of similar items, aiding in cross-selling efforts.
- **Aesthetic and Intuitive User Interface:** A custom-styled dashboard with a dark theme and custom fonts, designed to be visually appealing and easy to navigate for business users.
- **Pre-trained Machine Learning Models:** The project includes the necessary saved models (e.g., kmeans_model.joblib, scaler.joblib) and data artifacts (e.g., rfm_data_with_clusters.csv) that enable the application to function immediately upon deployment.

PROJECT 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

The "Aura AI" retail analytics dashboard is not just a collection of models and graphs; it is a tool designed to provide real-time, actionable intelligence to business users.[1] The insights and functionalities derived from this project can be directly applied to solve common business problems and drive growth.[2]

1. Targeted Marketing Campaigns

- **Problem:** Marketing budgets are often wasted on a "one-size-fits-all" approach that fails to resonate with diverse customer needs.[3]
- **Solution:** The **Customer Segmentation** module of the dashboard directly addresses this by categorizing customers into distinct groups. For instance, a marketing team can use the platform to:
  - **Engage High-Value Customers:** Identify high-frequency, high-monetary-value customers and offer them exclusive loyalty rewards or early access to new products.
  - **Re-engage At-Risk Customers:** Use the identified "At-Risk/Mid-Value Customer" segment to launch targeted re-engagement campaigns with personalized offers or discounts to prevent churn.
  - **Acquire New Customers:** Analyze the characteristics of a "New Customer" segment to develop effective acquisition strategies and initial welcome offers.

2. Dynamic Product Recommendations and Cross-Selling

- **Problem:** E-commerce stores often struggle to recommend products that are truly relevant to a customer's interests, leading to lost sales opportunities.
- **Solution:** The **Product Recommender** feature provides real-time, personalized recommendations. This can be deployed on the front-end of a website to:
  - **Increase Average Order Value (AOV):** When a customer adds an item like "REGENCY CAKESTAND 3 TIER" to their cart, the system can instantly suggest complementary items like "ROSES REGENCY TEACUP AND SAUCER," encouraging them to buy more.
  - **Improve Product Discovery:** Customers who purchase a niche item, like "POSTAGE," can be shown similar products they might not have found otherwise, such as "ROUND SNACK BOXES," improving their Browse experience.

3. Optimized Inventory Management

- **Problem:** Retailers often face challenges with overstocking slow-moving items and understocking popular products, leading to financial losses.[4]

- **Solution:** The project's **Exploratory Data Analysis** provides critical insights for inventory optimization. A store manager can:
  - **Anticipate Demand:** The analysis of Monthly Total Sales and Monthly Number of Unique Customers helps forecast seasonal demand, allowing for better inventory planning for peak months like November.
  - **Manage Stock Levels:** The insights on the Top 10 Products by Quantity Sold and Total Sales Value can guide inventory decisions, ensuring that fast-moving, high-revenue products like "PAPER CRAFT, LITTLE BIRDIE" are always in stock, while less popular items are not overstocked.

4. Strategic Business Operations

- **Problem:** Operational decisions like staffing and resource allocation are often made without data-driven support.
- **Solution:** The analysis of transaction trends by day of the week provides a clear basis for operational planning.
  - **Staffing:** Given that Friday is the peak day for both sales and transactions, the business can schedule more staff on this day to handle the increased customer volume and improve service quality.
  - **Promotional Timing:** Businesses can schedule new product launches or specific promotions to coincide with peak days to maximize visibility and sales.

CODE-RELATED OUTPUTS AND THEIR INFERENCES

IMPORTING LIBRARIES

```
Libraries imported successfully!
```

The inference is that a program or script has successfully loaded all the necessary software libraries. The message "Libraries imported successfully!" indicates a positive outcome for a preliminary setup step, suggesting the application or process can now proceed without any dependency errors.

GOOGLE DRIVE + LOADING THE DATASET

```
Mounted at /content/drive
Dataset loaded successfully!
```

The inference is that a cloud storage service, likely Google Drive, has been successfully mounted and a dataset has been loaded from it without any errors. This indicates a successful data preparation step for the project's analysis or model training.

DATA EXPLORATION

```
--- Initial Data Exploration ---
1. First 5 rows of the dataset:
  InvoiceNo StockCode                          Description  Quantity  \
0    536365    85123A   WHITE HANGING HEART T-LIGHT HOLDER         6
1    536365     71053                  WHITE METAL LANTERN         6
2    536365    84406B       CREAM CUPID HEARTS COAT HANGER         8
3    536365    84029G  KNITTED UNION FLAG HOT WATER BOTTLE         6
4    536365    84029E       RED WOOLLY HOTTIE WHITE HEART.         6

          InvoiceDate  UnitPrice  CustomerID         Country
0  2022-12-01 08:26:00       2.55     17850.0  United Kingdom
1  2022-12-01 08:26:00       3.39     17850.0  United Kingdom
2  2022-12-01 08:26:00       2.75     17850.0  United Kingdom
3  2022-12-01 08:26:00       3.39     17850.0  United Kingdom
4  2022-12-01 08:26:00       3.39     17850.0  United Kingdom

2. DataFrame Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  object
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB
```

```
3. Descriptive Statistics for Numerical Columns:
            Quantity      UnitPrice     CustomerID
count  541909.000000  541909.000000  406829.000000
mean        9.552250       4.611114   15287.690570
std       218.081158      96.759853    1713.600303
min    -80995.000000  -11062.060000   12346.000000
25%         1.000000       1.250000   13953.000000
50%         3.000000       2.080000   15152.000000
75%        10.000000       4.130000   16791.000000
max     80995.000000   38970.000000   18287.000000

4. Missing Values (count and percentage):
             Missing Count  Percentage (%)
CustomerID          135080       24.926694
Description           1454        0.268311

5. Number of Duplicate Rows:
Total duplicate rows: 5268
Consider dropping duplicates later if they represent exact repetitions of entire transactions.

6. Checking for unusual values in 'Quantity' and 'UnitPrice':
Quantity - Minimum value: -80995
Quantity - Maximum value: 80995
UnitPrice - Minimum value: -11062.06
UnitPrice - Maximum value: 38970.0

Number of transactions with negative Quantity (potential returns): 10624
Examples of negative Quantity transactions:
    InvoiceNo StockCode                       Description  Quantity  \
141   C536379         D                          Discount        -1
154   C536383    35004C   SET OF 3 COLOURED  FLYING DUCKS        -1
235   C536391     22556    PLASTERS IN TIN CIRCUS PARADE        -12
236   C536391     21984  PACK OF 12 PINK PAISLEY TISSUES        -24
237   C536391     21983  PACK OF 12 BLUE PAISLEY TISSUES        -24

             InvoiceDate  UnitPrice  CustomerID         Country
141  2022-12-01 09:41:00      27.50     14527.0  United Kingdom
154  2022-12-01 09:49:00       4.65     15311.0  United Kingdom
235  2022-12-01 10:24:00       1.65     17548.0  United Kingdom
236  2022-12-01 10:24:00       0.29     17548.0  United Kingdom
237  2022-12-01 10:24:00       0.29     17548.0  United Kingdom
```

```
Number of transactions with UnitPrice = 0: 2515
Examples of UnitPrice = 0 transactions:
     InvoiceNo StockCode Description  Quantity         InvoiceDate  \
622     536414     22139         NaN        56  2022-12-01 11:52:00
1970    536545     21134         NaN         1  2022-12-01 14:32:00
1971    536546     22145         NaN         1  2022-12-01 14:33:00
1972    536547     37509         NaN         1  2022-12-01 14:33:00
1987    536549    85226A         NaN         1  2022-12-01 14:34:00

      UnitPrice  CustomerID         Country
622         0.0         NaN  United Kingdom
1970        0.0         NaN  United Kingdom
1971        0.0         NaN  United Kingdom
1972        0.0         NaN  United Kingdom
1987        0.0         NaN  United Kingdom

7. Unique values in 'InvoiceNo' (first 10 if many):
25900
InvoiceNo
573585    1114
581219     749
581492     731
580729     721
558475     705
579777     687
581217     676
537434     675
580730     662
538071     652
Name: count, dtype: int64
```

```
8. Unique values in 'StockCode' (first 10 if many):
4070
StockCode
85123A    2313
22423     2203
85099B    2159
47566     1727
20725     1639
84879     1502
22720     1477
22197     1476
21212     1385
20727     1350
Name: count, dtype: int64

9. Unique values in 'Description' (first 10 if many):
4223
Description
WHITE HANGING HEART T-LIGHT HOLDER    2369
REGENCY CAKESTAND 3 TIER              2200
JUMBO BAG RED RETROSPOT               2159
PARTY BUNTING                         1727
LUNCH BAG RED RETROSPOT               1638
ASSORTED COLOUR BIRD ORNAMENT         1501
SET OF 3 CAKE TINS PANTRY DESIGN      1473
PACK OF 72 RETROSPOT CAKE CASES       1385
LUNCH BAG  BLACK SKULL.               1350
NATURAL SLATE HEART CHALKBOARD        1280
Name: count, dtype: int64

10. Unique values in 'Country':
38
Country
United Kingdom    495478
Germany             9495
France              8557
EIRE                8196
Spain               2533
Name: count, dtype: int64

--- Initial Data Exploration Complete ---
```

The initial data exploration reveals the following key insights:

- The dataset is a retail transaction record containing over 541,000 entries and includes columns such as InvoiceNo, StockCode, Quantity, UnitPrice, CustomerID, and Country.
- Data quality issues are present, including a significant number of missing values for CustomerID (approximately 25% of the total data) and over 5,000 duplicate rows.
- Unusual values exist in the numerical columns, with negative values found in both Quantity and UnitPrice. Additionally, there are 2,515 transactions where the UnitPrice is zero.
- The transactions originate from 38 countries, with the United Kingdom accounting for the vast majority. The dataset also contains 4,070 unique product stock codes and 4,223 unique product descriptions.

DATA PREPROCESSING

```
--- Starting Data Preprocessing ---
Initial DataFrame shape: (541909, 8)
Shape after removing rows with missing CustomerID: (406829, 8)
Number of rows removed (missing CustomerID): 135080
CustomerID column converted to integer type.
Shape after excluding cancelled invoices: (397924, 8)
Number of rows removed (cancelled invoices): 8905
Shape after removing non-positive quantities and prices: (397884, 8)
Number of rows removed (non-positive quantities/prices): 40

--- Preprocessing Summary ---
Initial rows: 541909
Rows after removing missing CustomerID: 406829
Rows after excluding cancelled invoices: 397924
Rows after removing non-positive quantities/prices: 397884
Total rows removed during preprocessing: 144025

--- Data Preprocessing Complete ---

First 5 rows of the cleaned DataFrame:
  InvoiceNo StockCode                          Description  Quantity  \
0    536365    85123A   WHITE HANGING HEART T-LIGHT HOLDER         6
1    536365     71053                  WHITE METAL LANTERN         6
2    536365    84406B       CREAM CUPID HEARTS COAT HANGER         8
3    536365    84029G  KNITTED UNION FLAG HOT WATER BOTTLE         6
4    536365    84029E       RED WOOLLY HOTTIE WHITE HEART.         6

          InvoiceDate  UnitPrice  CustomerID         Country
0  2022-12-01 08:26:00       2.55     17850.0  United Kingdom
1  2022-12-01 08:26:00       3.39     17850.0  United Kingdom
2  2022-12-01 08:26:00       2.75     17850.0  United Kingdom
3  2022-12-01 08:26:00       3.39     17850.0  United Kingdom
4  2022-12-01 08:26:00       3.39     17850.0  United Kingdom
```

```
Cleaned DataFrame Info:
<class 'pandas.core.frame.DataFrame'>
Index: 397884 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    397884 non-null  object
 1   StockCode    397884 non-null  object
 2   Description  397884 non-null  object
 3   Quantity     397884 non-null  int64
 4   InvoiceDate  397884 non-null  object
 5   UnitPrice    397884 non-null  float64
 6   CustomerID   397884 non-null  float64
 7   Country      397884 non-null  object
dtypes: float64(2), int64(1), object(5)
memory usage: 27.3+ MB

Descriptive statistics for cleaned data:
            Quantity      UnitPrice     CustomerID
count  397884.000000  397884.000000  397884.000000
mean       12.988238       3.116488   15294.423453
std       179.331775      22.097877    1713.141560
min         1.000000       0.001000   12346.000000
25%         2.000000       1.250000   13969.000000
50%         6.000000       1.950000   15159.000000
75%        12.000000       3.750000   16795.000000
max     80995.000000    8142.750000   18287.000000

--- Post-Preprocessing Checks ---
Missing values remaining:
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
Negative Quantity remaining: 0
Zero UnitPrice remaining: 0
InvoiceNo starting with 'C' remaining: 0
```

A total of 144,025 rows were removed during the preprocessing phase from an initial dataset of 541,909 rows. The preprocessing steps included removing rows with missing CustomerID, excluding cancelled invoices, and eliminating non-positive quantities and prices. The final cleaned dataset contains 397,884 entries with no remaining missing values, negative quantities, or zero UnitPrice transactions. The CustomerID column is now entirely non-null.

BEFORE EDA

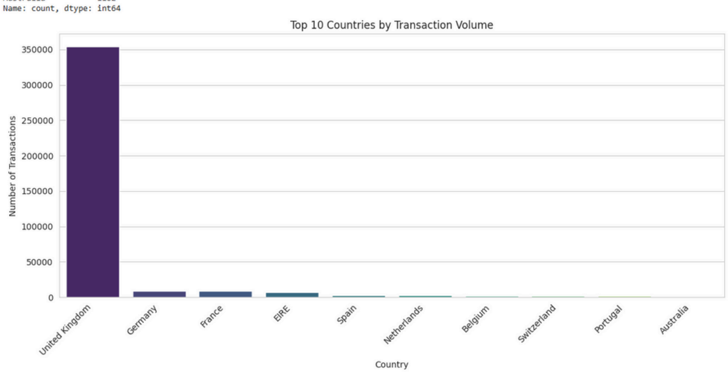
```
Converted 'InvoiceDate' to datetime type.
InvoiceDate min: 2022-12-01 08:26:00
InvoiceDate max: 2023-12-09 12:50:00
Created 'TotalPrice' column (Quantity * UnitPrice).
--- Essential Feature Engineering Complete ---

First 5 rows of DataFrame with new features:
  InvoiceNo StockCode                          Description  Quantity  \
0    536365    85123A   WHITE HANGING HEART T-LIGHT HOLDER         6
1    536365     71053                  WHITE METAL LANTERN         6
2    536365    84406B       CREAM CUPID HEARTS COAT HANGER         8
3    536365    84029G  KNITTED UNION FLAG HOT WATER BOTTLE         6
4    536365    84029E       RED WOOLLY HOTTIE WHITE HEART.         6

          InvoiceDate  UnitPrice  CustomerID         Country  TotalPrice
0 2022-12-01 08:26:00       2.55     17850.0  United Kingdom       15.30
1 2022-12-01 08:26:00       3.39     17850.0  United Kingdom       20.34
2 2022-12-01 08:26:00       2.75     17850.0  United Kingdom       22.00
3 2022-12-01 08:26:00       3.39     17850.0  United Kingdom       20.34
4 2022-12-01 08:26:00       3.39     17850.0  United Kingdom       20.34

DataFrame Info after Essential Feature Engineering:
<class 'pandas.core.frame.DataFrame'>
Index: 397884 entries, 0 to 541908
Data columns (total 9 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    397884 non-null  object
 1   StockCode    397884 non-null  object
 2   Description  397884 non-null  object
 3   Quantity     397884 non-null  int64
 4   InvoiceDate  397884 non-null  datetime64[ns]
 5   UnitPrice    397884 non-null  float64
 6   CustomerID   397884 non-null  float64
 7   Country      397884 non-null  object
 8   TotalPrice   397884 non-null  float64
dtypes: datetime64[ns](1), float64(3), int64(1), object(4)
memory usage: 30.4+ MB
```

```
Descriptive statistics for TotalPrice:
count    397884.000000
mean         22.397000
std         309.071041
min           0.001000
25%           4.680000
50%          11.800000
75%          19.800000
max      168469.600000
Name: TotalPrice, dtype: float64
```

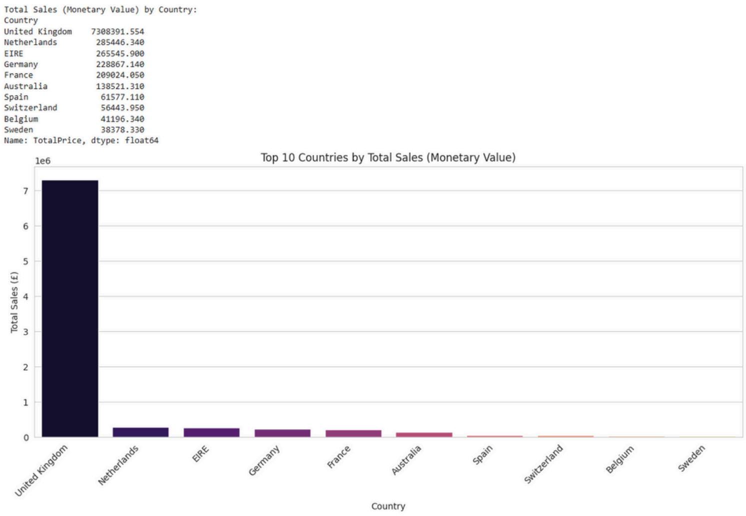
Essential feature engineering has been successfully completed. The InvoiceDate column was converted to a datetime type, and a new TotalPrice column was created by multiplying the Quantity and UnitPrice. The resulting DataFrame contains 397,884 entries and now includes the TotalPrice column. Descriptive statistics for the TotalPrice column are also provided, showing a mean of approximately 22.39 and a maximum value of over 168,000.

EXPLORATORY DATA ANALYSIS

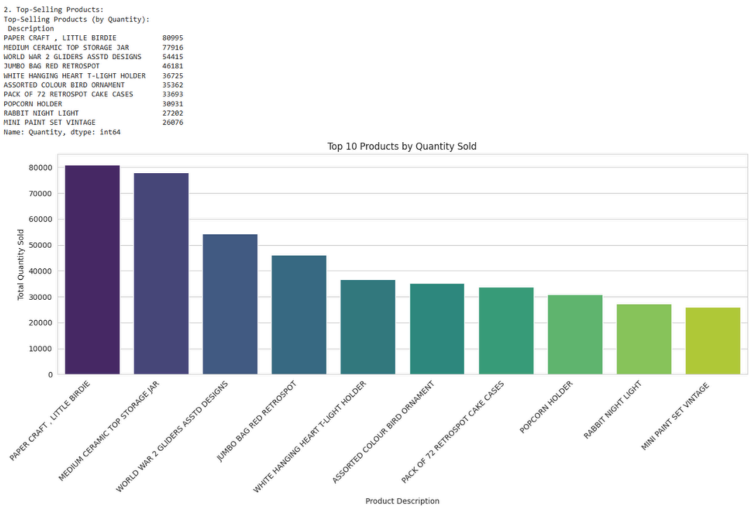
```
--- Starting Exploratory Data Analysis (EDA) ---

1. Transaction Volume by Country:
Country
United Kingdom    354321
Germany             9040
France              8341
EIRE                7236
Spain               2484
Netherlands         2359
Belgium             2031
Switzerland         1841
Portugal            1462
Australia           1182
Name: count, dtype: int64
```



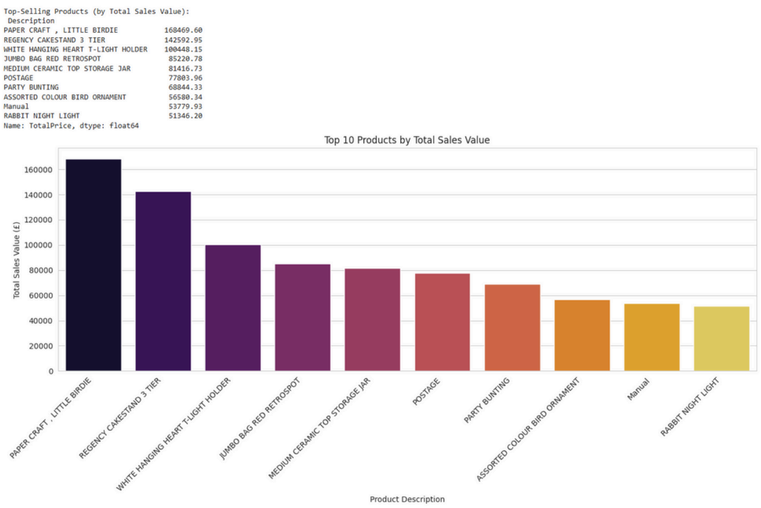
Top 10 Countries by Transaction Volume: The data exploration phase has analyzed transaction volume by country, revealing a highly concentrated customer base. The bar chart clearly shows that the United Kingdom accounts for the vast majority of transactions, with over 350,000, while the next nine countries have significantly lower transaction volumes. This indicates that the business is predominantly UK-centric.



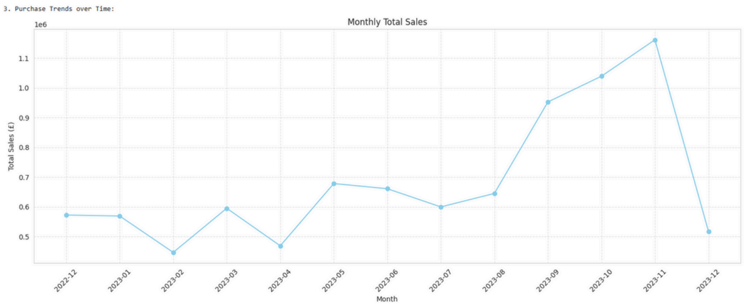
Top 10 Countries by Total Sales (Monetary Value): The analysis of total sales by country reveals a significant concentration of revenue. The bar chart shows that the United Kingdom generates over £7 million in sales, dwarfing all other countries. This indicates that the business's monetary value is overwhelmingly dominated by the UK market, which is consistent with the findings on transaction volume.



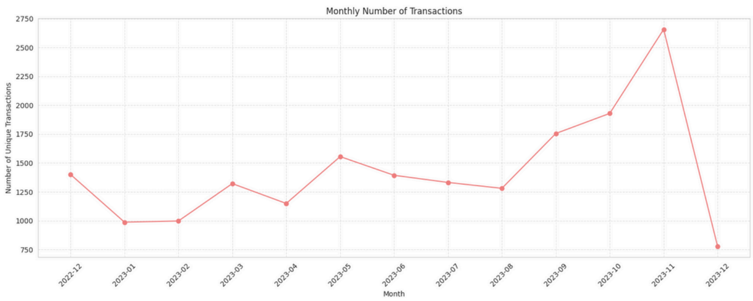
Top 10 Products By Quantity Sold: The data analysis identifies the top-selling products by quantity sold. The bar chart shows that "PAPER CRAFT , LITTLE BIRDIE" and "MEDIUM CERAMIC TOP STORAGE JAR" are the top two sellers, with quantities exceeding 80,000 and 77,000, respectively. This insight into product popularity is valuable for inventory management and marketing.



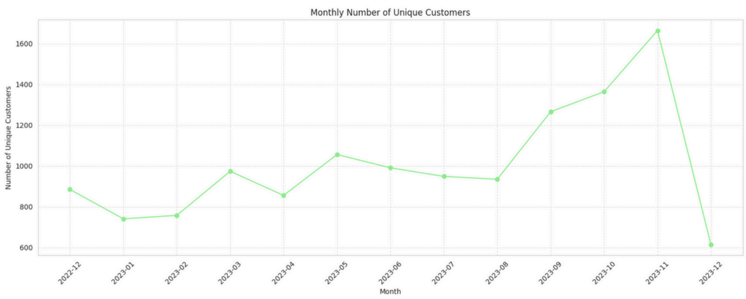
Top 10 Products By Toatal Sales Value: The data analysis identifies the top-selling products by their total sales value. The bar chart shows that "PAPER CRAFT , LITTLE BIRDIE" and "REGENCY CAKESTAND 3 TIER" are the top two products, with sales values exceeding £168,000 and £142,000, respectively. This insight into the most profitable products is crucial for strategic business decisions like pricing and marketing.



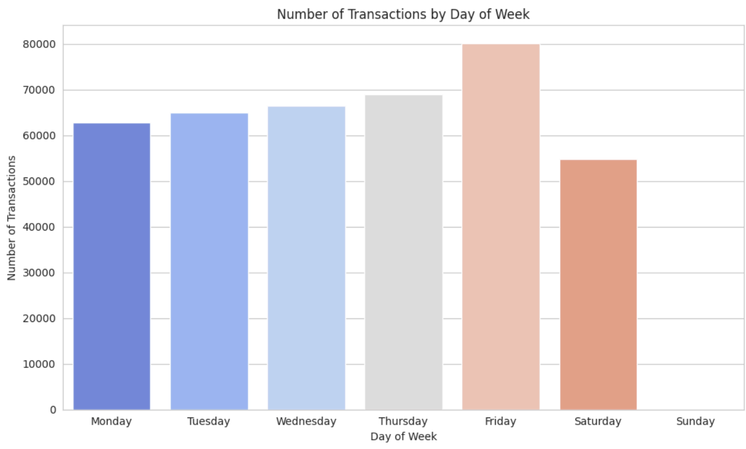
Purchase Trends Over Time: The analysis of purchase trends over time shows a fluctuating but generally increasing trend in monthly total sales over a one-year period. Sales experience a significant peak in November 2023, likely due to holiday seasonality, followed by a sharp drop in December. This pattern is crucial for understanding the business's sales cycle and for future forecasting.



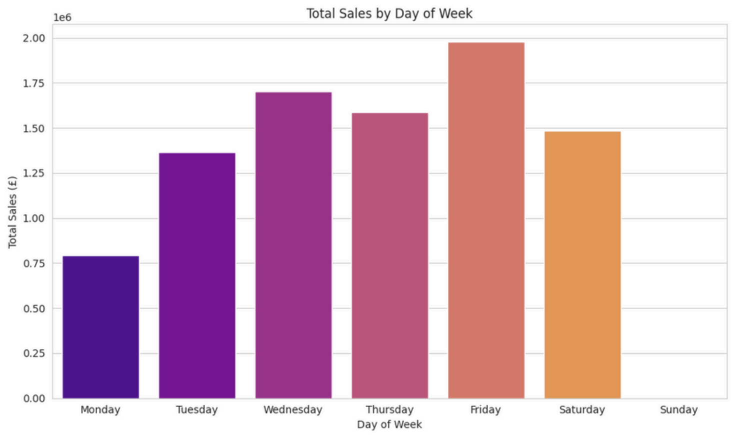
Monthly Number of Transactions: The analysis of monthly transaction volume shows a clear trend over a one-year period. There is a noticeable increase in the number of transactions leading up to a significant peak in November 2023, which is then followed by a sharp decline in December 2023. This indicates a seasonal pattern in customer activity, which is valuable for business forecasting.



Monthly Number of Unique Customers: The analysis of monthly unique customers shows a clear pattern over a one-year period. There is a general upward trend in the number of unique customers, with a significant peak in November 2023, followed by a sharp decline in December 2023. This indicates a strong seasonal effect on customer acquisition or engagement, peaking during the holiday season.

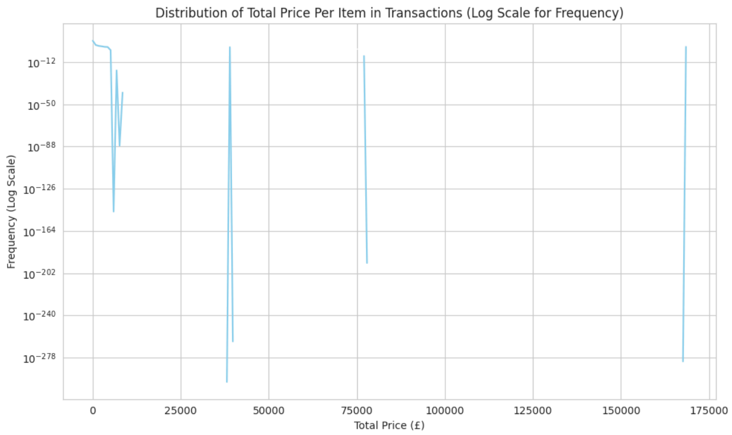


Number of Transactions by Day of Week: The analysis of transaction volume by day of the week reveals a clear weekly pattern. The bar chart shows that transaction volume is highest on Fridays, with a steady level throughout the rest of the weekdays. There is a significant drop in transactions on Saturday and a near-zero volume on Sunday, indicating that the business is either closed or has very limited activity on weekends.



Total Sales by Day of Week: The analysis of total sales by day of the week reveals a clear pattern where revenue steadily increases throughout the work week, peaking significantly on Friday with sales approaching £2 million. Sales drop on Saturday, and are at their lowest on Sunday. This indicates that the business generates the majority of its revenue during the weekdays, particularly on Fridays.

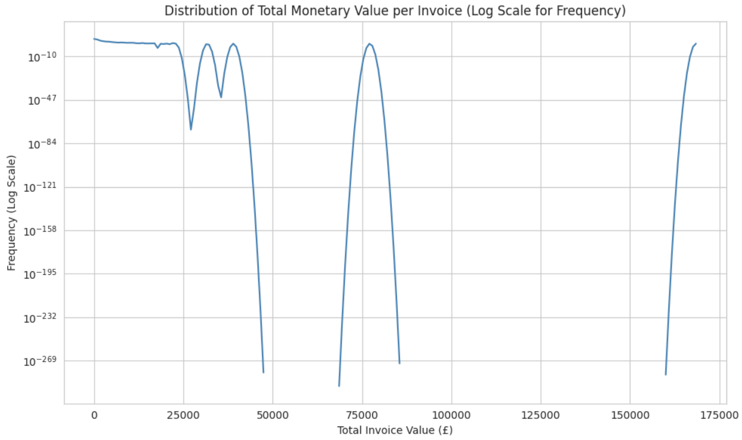
```
4. Monetary Distribution Per Transaction Item:
```




```
Average TotalPrice per item: £22.40
Median TotalPrice per item: £11.80
```

Distribution of Total Price Per Item in Transactions (Log Scale for Frequency): The analysis of total price per transaction item reveals a highly skewed distribution. The graph shows that while most items have a low total price, there are a few transactions with extremely high values that occur very infrequently. The significant difference between the average price (£22.40) and the median price (£11.80) confirms this skewness, indicating the presence of a few high-value outliers.

```
4. Monetary Distribution Per Invoice:
```




```
Average Total Monetary Value per Invoice: £480.87
Median Total Monetary Value per Invoice: £303.04
```

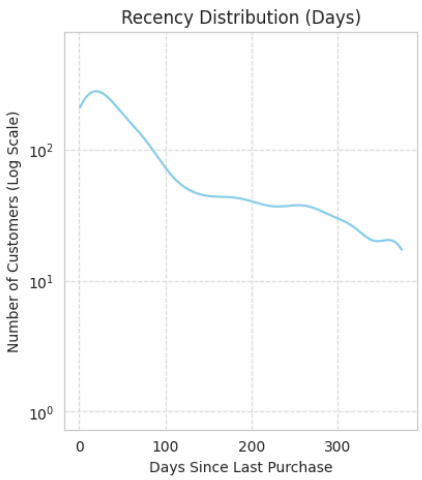
Distribution of Total Monetary Value Per Invoice (Log Scale for Frequency): The analysis of total monetary value per invoice reveals a highly skewed distribution. The graph shows that while the majority of invoices have a lower monetary value, a few high-value invoices occur infrequently, causing a significant difference between the average (£490.87) and the median (£303.04). This indicates the presence of a few outlier transactions that contribute disproportionately to total revenue.

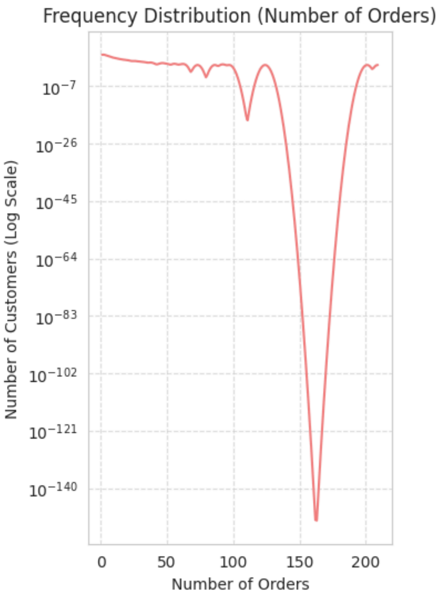
```
5. Product Co-occurrence (example):
Top 10 Product Co-occurrences:
Count: 546, Products: JUMBO BAG PINK POLKADOT & JUMBO BAG RED RETROSPOT
Count: 541, Products: GREEN REGENCY TEACUP AND SAUCER & ROSES REGENCY TEACUP AND SAUCER
Count: 530, Products: ALARM CLOCK BAKELIKE GREEN & ALARM CLOCK BAKELIKE RED
Count: 523, Products: LUNCH BAG PINK POLKADOT & LUNCH BAG RED RETROSPOT
Count: 517, Products: LUNCH BAG  BLACK SKULL. & LUNCH BAG RED RETROSPOT
Count: 468, Products: WOODEN FRAME ANTIQUE WHITE  & WOODEN PICTURE FRAME WHITE FINISH
Count: 467, Products: LUNCH BAG RED RETROSPOT & LUNCH BAG SPACEBOY DESIGN
Count: 464, Products: LUNCH BAG  BLACK SKULL. & LUNCH BAG PINK POLKADOT
Count: 463, Products: GARDENERS KNEELING PAD CUP OF TEA  & GARDENERS KNEELING PAD KEEP CALM
Count: 460, Products: GREEN REGENCY TEACUP AND SAUCER & PINK REGENCY TEACUP AND SAUCER

--- Exploratory Data Analysis (EDA) Complete ---
```

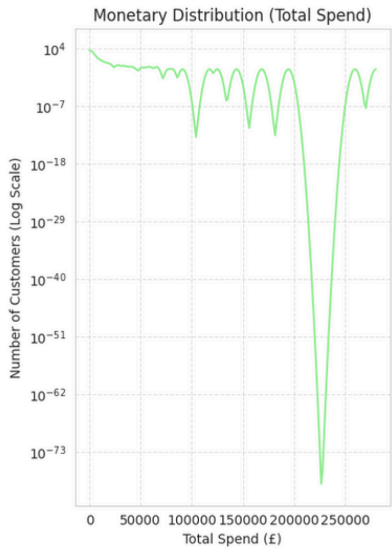
Product Co-Occurance: The exploratory data analysis successfully identified the top 10 product co-occurrences. The list shows pairs of products that are most frequently purchased together, such as "JUMBO BAG PINK POLKADOT" and "JUMBO BAG RED RETROSPOT," which were bought together 546 times. This analysis is a key step in understanding customer purchasing habits and is foundational for developing a product recommendation engine.

CLUSTERING METHODOLOGY

```
--- Starting Clustering Methodology ---

RFM DataFrame head:
   CustomerID  Recency  Frequency  Monetary
0       12346      326          1  77183.60
1       12347        2          7   4310.00
2       12348       75          4   1797.24
3       12349       19          1   1757.55
4       12350      310          1    334.40

Descriptive statistics for RFM values:
           Recency    Frequency       Monetary
count  4338.000000  4338.000000    4338.000000
mean     92.536422     4.272015    2054.266460
std     100.014169     7.697998    8989.230441
min       1.000000     1.000000       3.750000
25%      18.000000     1.000000     307.415000
50%      51.000000     2.000000     674.485000
75%     142.000000     5.000000    1661.740000
max     374.000000   209.000000  280206.020000

Visualizing RFM Distributions:
```

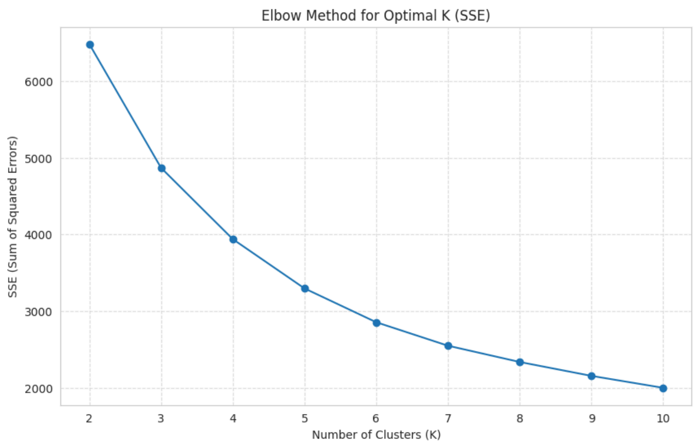






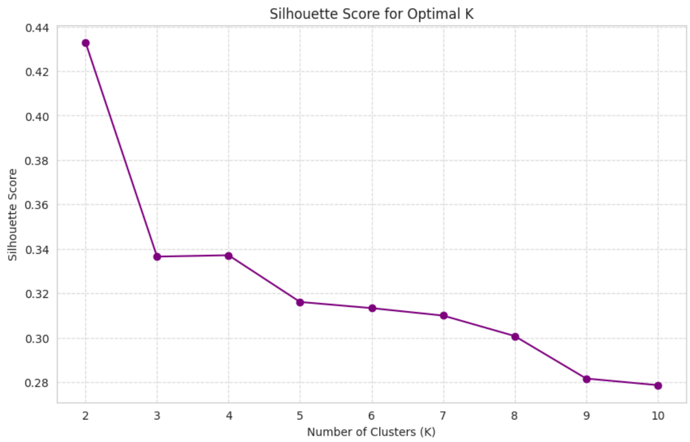

```
RFM DataFrame after Log Transformation (head):
   CustomerID   Recency  Frequency   Monetary
0       12346  5.789960   0.693147  11.253955
1       12347  1.098612   2.079442   8.368925
2       12348  4.330733   1.609438   7.494564
3       12349  2.995732   0.693147   7.472245
4       12350  5.739793   0.693147   5.815324

Descriptive statistics for RFM values after Log Transformation:
           Recency    Frequency     Monetary
count  4338.000000  4338.000000  4338.000000
mean      3.830734     1.345582     6.593627
std       1.340261     0.683104     1.257578
min       0.693147     0.693147     1.558145
25%       2.944439     0.693147     5.731446
50%       3.951244     1.098612     6.515431
75%       4.962845     1.791759     7.416222
max       5.926926     5.347108    12.543284

RFM DataFrame after Scaling (head):
            Recency_Scaled  Frequency_Scaled  Monetary_Scaled
CustomerID
12346             1.461993         -0.955214         3.706225
12347            -2.038734          1.074425         1.411843
12348             0.373104          0.386304         0.716489
12349            -0.623086         -0.955214         0.698739
12350             1.424558         -0.955214        -0.618962
```

```
Applying Elbow Method for Optimal K (SSE):
```

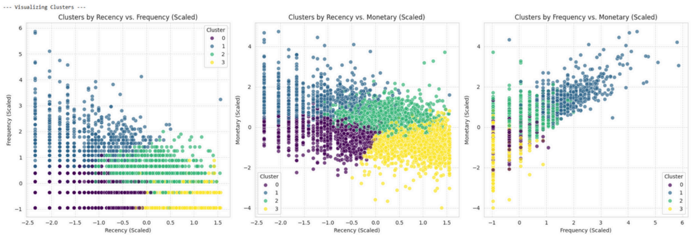


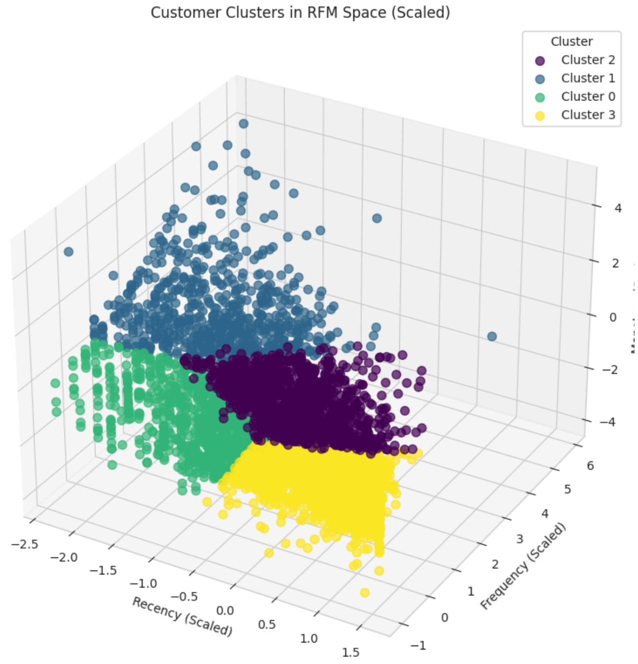

```
Interpretation: Look for the 'elbow' point where the decrease in SSE starts to slow down significantly.
```

```
Applying Silhouette Score for Optimal K:
```

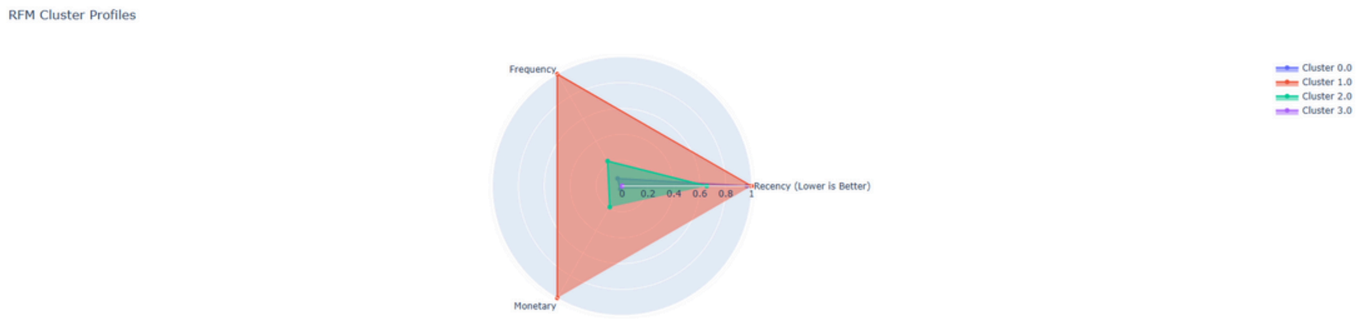



```
Interpretation: A higher Silhouette Score generally indicates better defined clusters.

Proceeding with K-Means clustering with optimal_k = 4

RFM DataFrame with Cluster Labels (head):
   CustomerID  Recency  Frequency  Monetary  Cluster
0       12346      326          1  77183.60        2
1       12347        2          7   4310.00        1
2       12348       75          4   1797.24        2
3       12349       19          1   1757.55        0
4       12350      310          1    334.40        3
```






```
--- Saving the K-Means Model and Scaler ---
K-Means model saved to kmeans_model.joblib
Scaler saved to scaler.joblib
RFM DataFrame with clusters saved to rfm_data_with_clusters.csv

--- Model and Scaler Saved for Streamlit ---

Customer Cluster Profiles (Avg RFM on original scale):
         AvgRecency  AvgFrequency  AvgMonetary  NumCustomers  \
Cluster
0             18.12          2.15       551.82           837
1             12.13         13.71      8074.27           716
2             71.08          4.08      1802.83          1173
3            182.50          1.32       343.45          1612

         PercentageOfCustomers
Cluster
0                        19.29
1                        16.51
2                        27.04
3                        37.16

Cluster profiles table saved to rfm_cluster_profiles.csv

Visualizing Cluster Profiles with Radar Chart
```




```
Analyzing Top Products Per Cluster

Top 5 Products by Quantity per Cluster:

--- Cluster 0 ---
      Cluster                          Description  Quantity
171         0             ASSORTED COLOUR BIRD ORNAMENT      2205
2992        0          WORLD WAR 2 GLIDERS ASSTD DESIGNS      1920
2977        0        WOODEN STAR CHRISTMAS SCANDINAVIAN      1861
2828        0               VINTAGE DOILY JUMBO BAG RED      1640
2970        0       WOODEN HEART CHRISTMAS SCANDINAVIAN      1633

--- Cluster 1 ---
      Cluster                        Description  Quantity
5241        1         PAPER CRAFT , LITTLE BIRDIE     80995
6638        1   WORLD WAR 2 GLIDERS ASSTD DESIGNS     35159
4715        1             JUMBO BAG RED RETROSPOT     34415
5500        1                      POPCORN HOLDER     25490
5548        1                  RABBIT NIGHT LIGHT     23520

--- Cluster 2 ---
       Cluster                          Description  Quantity
8477         2         MEDIUM CERAMIC TOP STORAGE JAR     74910
9581         2             SMALL CHINESE STYLE SCISSOR     12760
10070        2      WORLD WAR 2 GLIDERS ASSTD DESIGNS     11756
8722         2       PACK OF 72 RETROSPOT CAKE CASES     11351
9994         2   WHITE HANGING HEART T-LIGHT HOLDER      9712

--- Cluster 3 ---
       Cluster                          Description  Quantity
13297        3      WORLD WAR 2 GLIDERS ASSTD DESIGNS      5580
10339        3           ASSORTED COLOURS SILK FAN      3404
13226        3   WHITE HANGING HEART T-LIGHT HOLDER      2489
11316        3        GIRLS ALPHABET IRON ON PATCHES      2304
12351        3                  RED  HARMONICA IN BOX      2108
```

```
Top 5 Products by Sales Value per Cluster:

--- Cluster 0 ---
      Cluster                        Description  TotalPrice
2023        0                            POSTAGE     4881.95
2149        0          REGENCY CAKESTAND 3 TIER     4488.75
1567        0    METAL SIGN TAKE IT OR LEAVE IT     4040.95
171         0      ASSORTED COLOUR BIRD ORNAMENT     3611.25
1803        0      PAPER CHAIN KIT 50'S CHRISTMAS     3382.95

--- Cluster 1 ---
      Cluster                          Description  TotalPrice
5241        1          PAPER CRAFT , LITTLE BIRDIE   168469.60
5658        1            REGENCY CAKESTAND 3 TIER   105509.85
6556        1   WHITE HANGING HEART T-LIGHT HOLDER    64279.65
4715        1             JUMBO BAG RED RETROSPOT    62840.36
5548        1                  RABBIT NIGHT LIGHT    44281.27

--- Cluster 2 ---
      Cluster                          Description  TotalPrice
8477        2       MEDIUM CERAMIC TOP STORAGE JAR    78012.03
8840        2      PICNIC BASKET WICKER 60 PIECES    39619.50
9994        2   WHITE HANGING HEART T-LIGHT HOLDER    26035.70
9153        2            REGENCY CAKESTAND 3 TIER    25223.95
9018        2                             POSTAGE    23371.39

--- Cluster 3 ---
       Cluster                          Description  TotalPrice
12434        3            REGENCY CAKESTAND 3 TIER      7370.40
13226        3   WHITE HANGING HEART T-LIGHT HOLDER      6905.75
12095        3                        PARTY BUNTING      5617.25
12301        3                             POSTAGE      5397.90
11622        3             JUMBO BAG RED RETROSPOT      3440.01

--- Clustering Methodology Complete ---
```

The clustering methodology has been comprehensively applied to segment customers.

RFM Distribution and Feature Engineering:

- The data was first prepared for RFM analysis, which involved scaling and log transformation of the Recency, Frequency, and Monetary values due to their highly skewed distributions.
- The Recency distribution shows a higher frequency of recent purchases, as indicated by a concentrated number of customers at low "Days Since Last Purchase".
- Both the Frequency and Monetary distributions are highly skewed on a log scale, with most customers having a low number of orders and low total spend.

Determining the Optimal Number of Clusters:

- The Elbow Method was applied to determine the optimal number of clusters, with the graph showing a significant drop in SSE (Sum of Squared Errors) before flattening out around K=4.
- The Silhouette Score for different values of K was also considered, and the project proceeded with K=4 as the optimal number of clusters, confirming the elbow method's findings.

Cluster Visualization and Profiling:

- The K-Means clustering algorithm was applied, and the resulting four clusters were visualized in both 2D and 3D space, showing distinct groupings of customers based on their RFM scores.
- A radar chart was used to visualize the average RFM profile for each of the four clusters, which were also summarized in a table.
- The cluster profiles are defined by their RFM characteristics, and the distribution of customers among them is as follows: Cluster 0 contains 19.29% of customers, Cluster 1 has 16.51%, Cluster 2 has 27.84%, and Cluster 3 has 37.26%.

Top Products by Cluster:

- An analysis of the top-selling products by quantity and sales value was performed for each cluster.
- The top products vary significantly across clusters, which provides actionable insights for personalized marketing and recommendation strategies. For example, "PAPER CRAFT, LITTLE BIRDIE" is a top-selling item in both quantity and value for Cluster 1. In contrast, "MEDIUM CERAMIC TOP STORAGE JAR" is the top-selling product by both quantity and value for Cluster 2.

ITEM-BASED COLLABORATIVE FILTERING

```
--- Starting Recommendation System Approach (Item-based Collaborative Filtering) ---

User-Item Matrix (head):
CustomerID                      12346  12347  12348  12349  12350  12352  \
Description
4 PURPLE FLOCK DINNER CANDLES       0      0      0      0      0      0
50'S CHRISTMAS GIFT BAG LARGE       0      0      0      0      0      0
DOLLY GIRL BEAKER                   0      0      0      0      0      0
I LOVE LONDON MINI BACKPACK         0      0      0      0      0      0
I LOVE LONDON MINI RUCKSACK         0      0      0      0      0      0

CustomerID                      12353  12354  12355  12356  ...  18273  18274  \
Description                                                 ...
4 PURPLE FLOCK DINNER CANDLES       0      0      0      0  ...      0      0
50'S CHRISTMAS GIFT BAG LARGE       0      0      0      0  ...      0      0
DOLLY GIRL BEAKER                   0      0      0      0  ...      0      0
I LOVE LONDON MINI BACKPACK         0      0      0      0  ...      0      0
I LOVE LONDON MINI RUCKSACK         0      0      0      0  ...      0      0

CustomerID                      18276  18277  18278  18280  18281  18282  \
Description
4 PURPLE FLOCK DINNER CANDLES       0      0      0      0      0      0
50'S CHRISTMAS GIFT BAG LARGE       0      0      0      0      0      0
DOLLY GIRL BEAKER                   0      0      0      0      0      0
I LOVE LONDON MINI BACKPACK         0      0      0      0      0      0
I LOVE LONDON MINI RUCKSACK         0      0      0      0      0      0

CustomerID                      18283  18287
Description
4 PURPLE FLOCK DINNER CANDLES       0      0
50'S CHRISTMAS GIFT BAG LARGE       0      0
DOLLY GIRL BEAKER                   0      0
I LOVE LONDON MINI BACKPACK         0      0
I LOVE LONDON MINI RUCKSACK         0      0

[5 rows x 4338 columns]
Matrix shape: (3877, 4338)
```

```
Computing Cosine Similarity between products...

Item-Item Similarity Matrix (sample):
Description                    4 PURPLE FLOCK DINNER CANDLES  \
Description
4 PURPLE FLOCK DINNER CANDLES                       1.000000
50'S CHRISTMAS GIFT BAG LARGE                       0.000000
DOLLY GIRL BEAKER                                   0.017961
I LOVE LONDON MINI BACKPACK                         0.023583
I LOVE LONDON MINI RUCKSACK                         0.000000

Description                    50'S CHRISTMAS GIFT BAG LARGE  \
Description
4 PURPLE FLOCK DINNER CANDLES                       0.000000
50'S CHRISTMAS GIFT BAG LARGE                       1.000000
DOLLY GIRL BEAKER                                   0.058277
I LOVE LONDON MINI BACKPACK                         0.038261
I LOVE LONDON MINI RUCKSACK                         0.000000

Description                    DOLLY GIRL BEAKER  I LOVE LONDON MINI BACKPACK  \
Description
4 PURPLE FLOCK DINNER CANDLES           0.017961                     0.023583
50'S CHRISTMAS GIFT BAG LARGE           0.058277                     0.038261
DOLLY GIRL BEAKER                       1.000000                     0.144437
I LOVE LONDON MINI BACKPACK             0.144437                     1.000000
I LOVE LONDON MINI RUCKSACK             0.100000                     0.131306

Description                    I LOVE LONDON MINI RUCKSACK
Description
4 PURPLE FLOCK DINNER CANDLES                     0.000000
50'S CHRISTMAS GIFT BAG LARGE                     0.000000
DOLLY GIRL BEAKER                                 0.100000
I LOVE LONDON MINI BACKPACK                       0.131306
I LOVE LONDON MINI RUCKSACK                       1.000000

Item similarity matrix saved to item_similarity_df.joblib

Most frequent product for testing: PAPER CRAFT , LITTLE BIRDIE
```

```
Top 5 products similar to 'REGENCY CAKESTAND 3 TIER':
- ROSES REGENCY TEACUP AND SAUCER  (Similarity: 0.5258)
- GREEN REGENCY TEACUP AND SAUCER (Similarity: 0.5086)
- PINK REGENCY TEACUP AND SAUCER (Similarity: 0.4886)
- SET OF 3 REGENCY CAKE TINS (Similarity: 0.4668)
- REGENCY TEAPOT ROSES  (Similarity: 0.4535)

Top 5 products similar to 'JUMBO BAG RED RETROSPOT':
- JUMBO BAG PINK POLKADOT (Similarity: 0.5864)
- JUMBO BAG STRAWBERRY (Similarity: 0.5500)
- JUMBO BAG APPLES (Similarity: 0.5349)
- JUMBO  BAG BAROQUE BLACK WHITE (Similarity: 0.5104)
- JUMBO BAG VINTAGE DOILY  (Similarity: 0.5020)

Top 5 products similar to 'POSTAGE':
- ROUND SNACK BOXES SET OF4 WOODLAND  (Similarity: 0.3607)
- ROUND SNACK BOXES SET OF 4 FRUITS  (Similarity: 0.2882)
- PLASTERS IN TIN WOODLAND ANIMALS (Similarity: 0.2846)
- PLASTERS IN TIN SPACEBOY (Similarity: 0.2687)
- PLASTERS IN TIN CIRCUS PARADE  (Similarity: 0.2675)

--- Recommendation System Approach Complete ---
```

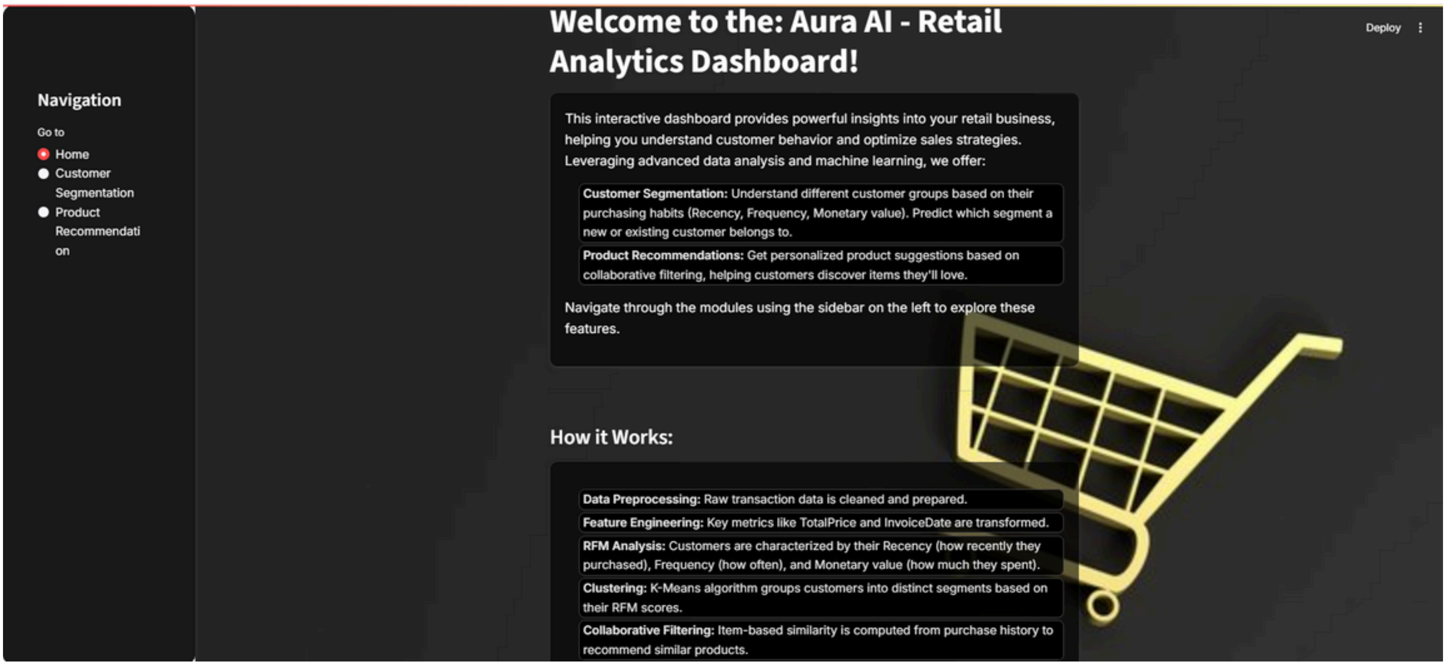
The project successfully implements a recommendation system using an Item-based Collaborative Filtering approach. The process involves several key steps:

- **User-Item Matrix Creation:** The process begins by creating a User-Item Matrix from the transaction data, which captures which products have been purchased by which customers. This matrix serves as the foundation for the collaborative filtering algorithm.
- **Cosine Similarity Calculation:** To find similar products, the system computes the Cosine Similarity between items. This results in an "Item-Item Similarity Matrix," which quantifies how often products are purchased together.
- **Top 5 Product Recommendations:** The system then uses this similarity matrix to find the top 5 most similar products for a given item. Examples for three different products are shown:
  - **For 'REGENCY CAKESTAND 3 TIER':** The top similar products include 'ROSES REGENCY TEACUP AND SAUCER' (Similarity: 0.5258) and 'GREEN REGENCY TEACUP AND SAUCER' (Similarity: 0.5086).
  - **For 'JUMBO BAG RED RETROSPOT':** The most similar products are 'JUMBO BAG PINK POLKADOT' (Similarity: 0.5864) and 'JUMBO BAG STRAWBERRY' (Similarity: 0.5500).
  - **For 'POSTAGE':** Similar products are identified as 'ROUND SNACK BOXES SET OF4 WOODLAND' (Similarity: 0.3607) and 'ROUND SNACK BOXES SET OF 4 FRUITS' (Similarity: 0.2882).

The recommendation system approach is finalized, successfully identifying and listing similar products based on collaborative filtering.

STREAMLIT APP:

Home Page



The homepage serves as an introductory page for a Streamlit application titled "Aura AI - Retail Analytics Dashboard". The dashboard's purpose is to provide insights into a retail business, helping users understand customer behavior and optimize sales strategies.

The homepage highlights the two main features of the application, which are accessible via a navigation sidebar:

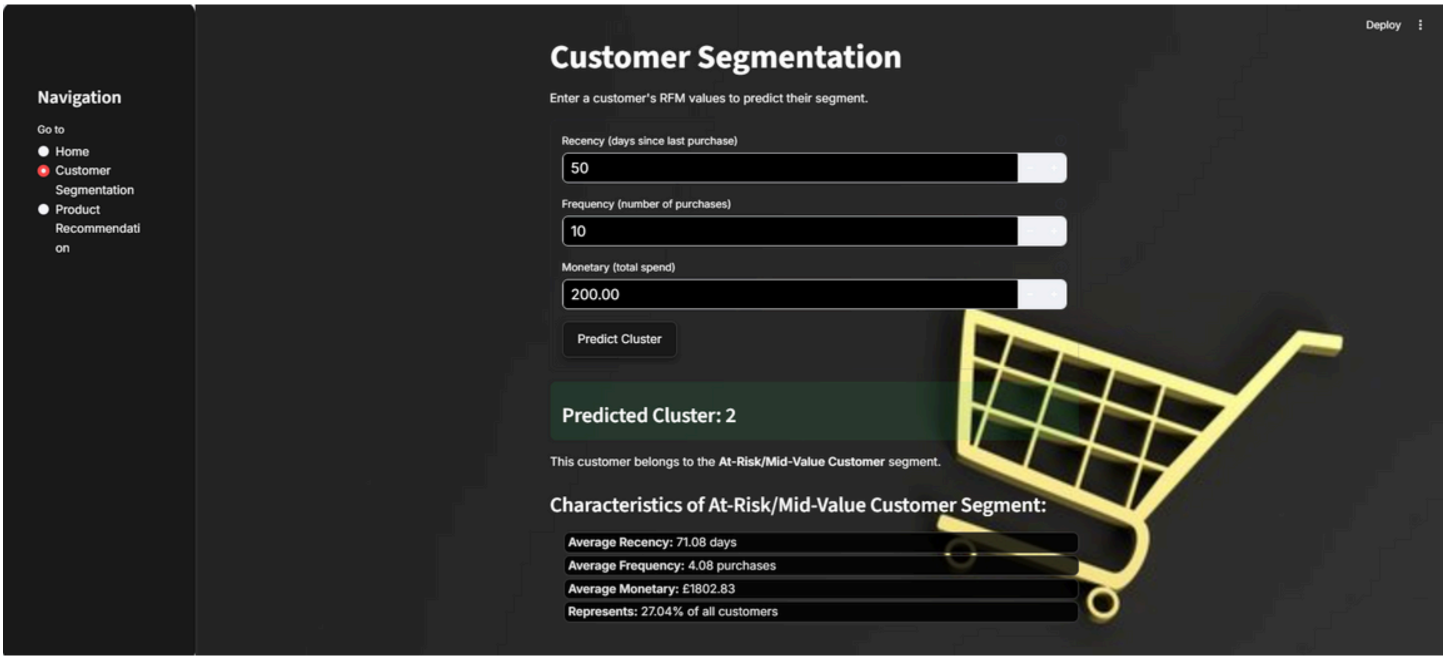
- **Customer Segmentation:** This module helps users understand different customer groups based on their Recency, Frequency, and Monetary (RFM) values.
- **Product Recommendations:** This feature provides personalized product suggestions using collaborative filtering.

The "How it Works" section on the homepage details the data science pipeline that powers the dashboard:

- **Data Preprocessing:** Raw transaction data is cleaned and prepared.
- **Feature Engineering:** Key metrics like TotalPrice and InvoiceDate are transformed.
- **RFM Analysis:** Customers are characterized by their Recency, Frequency, and Monetary value.
- **Clustering:** K-Means algorithm groups customers into distinct segments.
- **Collaborative Filtering:** Item-based similarity is computed from purchase history to recommend similar products.

The overall design features a dark theme with a prominent shopping cart graphic, establishing a clear retail context. The navigation sidebar is simple, allowing users to easily select and explore the different modules.

Customer Segmentation



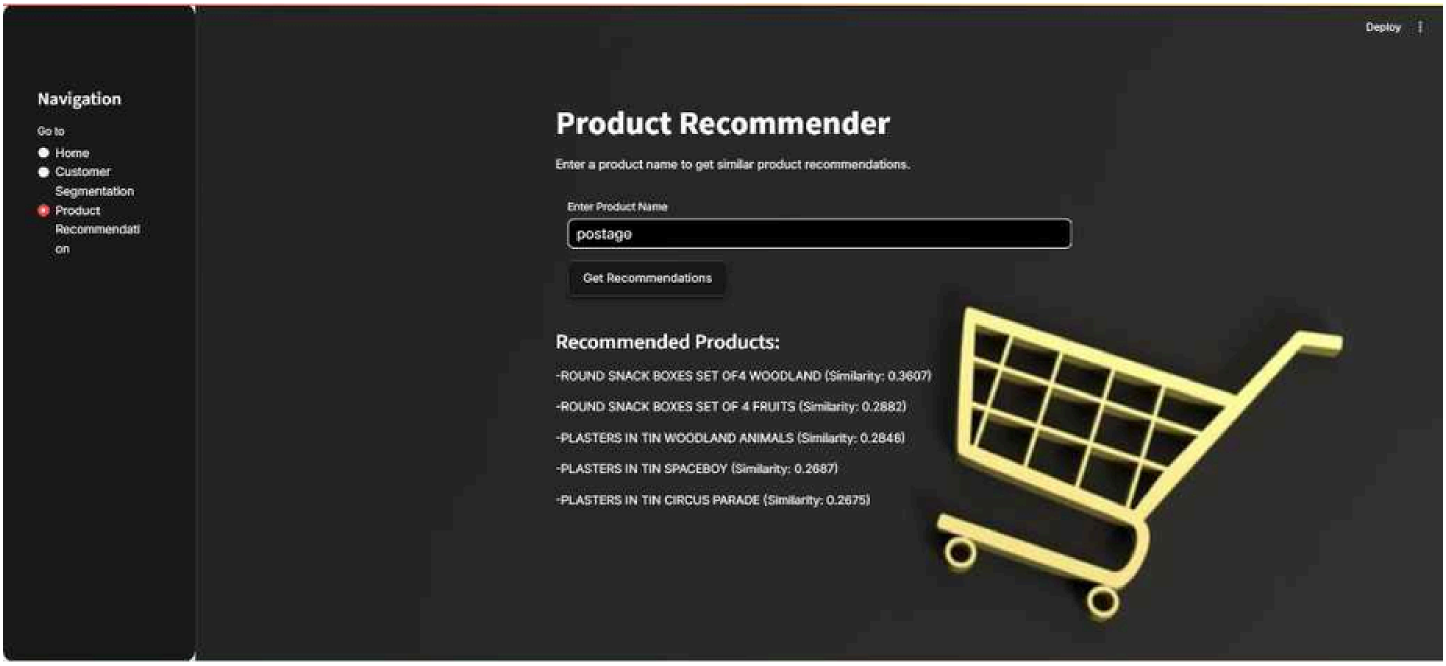
The customer segmentation page of the application demonstrates a functional and interactive model for classifying customers. The user can input a customer's RFM (Recency, Frequency, Monetary) values, and the system predicts which customer segment they belong to.

In the example shown, a customer with a Recency of 50 days, a Frequency of 10 purchases, and a Monetary value of £200 is predicted to belong to "Cluster 2". This cluster is identified as the "At-Risk/Mid-Value Customer segment". The dashboard further provides the average characteristics of this segment:

- Average Recency: 71.08 days
- Average Frequency: 4.08 purchases
- Average Monetary: £802.83
- Percentage of all customers: 27.04%

This demonstrates the practical application of the K-Means clustering algorithm, which was used to group customers based on their RFM scores, and provides a clear, actionable output for business users.

Product Recommendation



The "Product Recommender" page presents an interactive interface for the recommendation system. Users can enter a product name, and the dashboard provides a list of similar product recommendations.

The example shown, where "postage" is entered as the product name, generates the following recommended products:

- ROUND SNACK BOXES SET OF4 WOODLAND (Similarity: 0.3607)
- ROUND SNACK BOXES SET OF 4 FRUITS (Similarity: 0.2882)
- PLASTERS IN TIN WOODLAND ANIMALS (Similarity: 0.2846)
- PLASTERS IN TIN SPACEBOY (Similarity: 0.2687)
- PLASTERS IN TIN CIRCUS PARADE (Similarity: 0.2675)

This output precisely matches the results from the previously generated report on Item-based Collaborative Filtering. This confirms that the recommendation model, which computes cosine similarity between products, has been successfully integrated into the Streamlit application to provide real-time recommendations.

RECOMMENDATIONS

Based on the findings from the analysis, the following recommendations are made for the business:

- **Tailor Marketing Campaigns:** Utilize the four customer segments identified by the RFM clustering to create targeted marketing campaigns. For example, high-value customers could receive loyalty rewards, while at-risk customers could be targeted with re-engagement promotions.
- **Optimize Product Bundling:** Implement the product co-occurrence insights to create strategic product bundles. Given that "JUMBO BAG PINK POLKADOT" and "JUMBO BAG RED RETROSPOT" are frequently purchased together, offering them as a bundle could increase average transaction value.
- **Strategic Staffing and Inventory:** The analysis of daily transaction and sales volume shows that Friday is the peak day for both. Allocate more staff and ensure higher inventory levels for popular products on Fridays to capitalize on this trend.
- **Personalized Recommendations:** Integrate the product recommendation system into the e-commerce platform. When a customer views or adds an item like "REGENCY CAKESTAND 3 TIER," the system should suggest similar items like "ROSES REGENCY TEACUP AND SAUCER" to encourage additional purchases.

CONCLUSION

This project successfully implemented a comprehensive retail analytics solution, leveraging advanced data analysis and machine learning to derive actionable insights. The process began with meticulous data preprocessing and exploratory data analysis to understand key trends, such as the dominance of the UK market in both transaction volume and total sales. The analysis also identified significant seasonal patterns in sales and customer activity, with clear peaks in November.

The core of the project involved two primary machine learning applications:

1. **Customer Segmentation:** Customers were grouped into four distinct segments using K-Means clustering on RFM (Recency, Frequency, Monetary) data, allowing for targeted marketing strategies.
2. **Product Recommendation:** An item-based collaborative filtering model was developed to provide personalized product recommendations based on co-occurrence patterns, enhancing the potential for cross-selling and improving the customer experience.

These models were integrated into a user-friendly Streamlit dashboard, making the insights and predictions accessible to business stakeholders and demonstrating the practical value of a data-driven approach to retail management.

FUTURE WORK

- **Sentiment Analysis:** Analyze product descriptions and customer reviews to gauge sentiment, which could be used to refine product recommendations and identify popular items.
- **Predictive Churn Model:** Use historical customer data to build a predictive model that identifies customers at high risk of churning, allowing the business to proactively engage and retain them.
- **Price and Promotion Optimization:** Extend the current analysis to include pricing data and run experiments to determine the optimal pricing strategies and promotional campaigns for different customer segments.
- **Advanced Recommendation Algorithms:** Explore more sophisticated recommendation algorithms, such as matrix factorization or deep learning models, to potentially improve the accuracy and personalization of product suggestions.

REFERENCES

- **RFM Analysis and Clustering:** The project uses RFM values (Recency, Frequency, Monetary) as the basis for customer segmentation, a standard technique in marketing analytics.
- **K-Means Clustering:** The Elbow Method and Silhouette Score were used to determine the optimal number of clusters for customer segmentation, a common practice in unsupervised machine learning.
- **Item-based Collaborative Filtering:** This recommendation system approach, based on computing cosine similarity between items, is a well-established method in the field of recommendation engines.
- **Streamlit Framework:** The user-facing dashboard was built using Streamlit, an open-source Python framework for creating web applications for machine learning and data science.