

AIGC-5503



MIDTERM



Daniel Mehta (n01753264)

July 3rd 2025





GOAL

Apply machine learning to extract actionable insights from retail transaction data and guide business decisions.

TASK #1

Market Basket Analysis

(Apriori Algorithm)

TASK #2

Customer Segmentation

(RFM + K-Means)

TASK #3

Country-Level Clustering

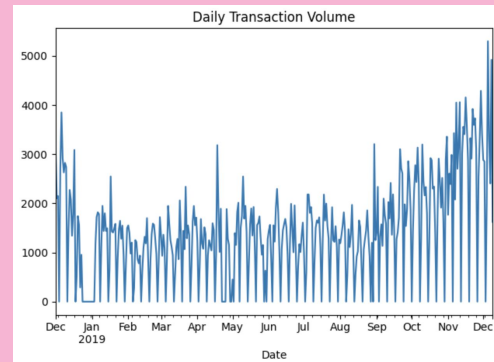
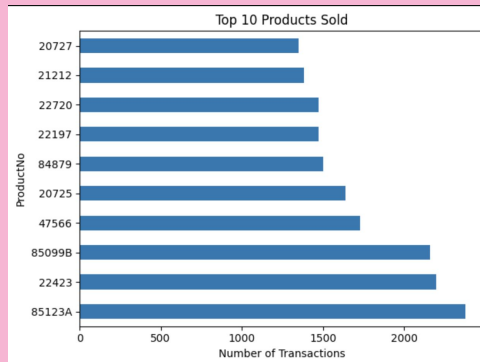
(Revenue, Activity via K-Means)



Dataset Overview

Customer, Product & Transaction Data

This project uses over 536,000 transaction records to explore customer behavior, product trends, and sales activity using integrated data from three sources: Transactions, Customers, and Products.



CUSTOMERS

Country, gender, and age metadata used for segmentation and grouping.

PRODUCTS

Each product includes a unique ID, price, and category

CLEANING

Removed rows with missing values in key columns, converted data types (dates, IDs), and merged the three datasets into one analytics-ready table.

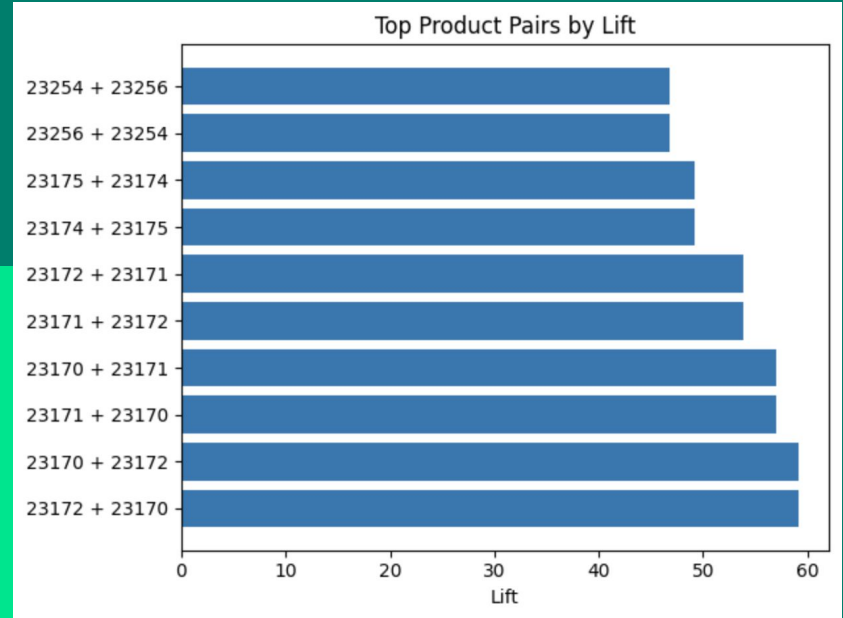
MARKET BASKET ANALYSIS (Apriori)

The goal of this task was to identify which products are frequently purchased together to support cross-selling strategies.

Using transaction-level purchase data, we applied the Apriori algorithm to uncover strong product associations based on support, confidence, and lift metrics.

Only rules with a minimum confidence of 40% and lift above 1.2 were retained to ensure relevance.

I chose Apriori due to its ability to efficiently identify frequent item combinations and generate interpretable association rules that support cross-selling strategies.



MARKET BASKET ANALYSIS:

INSIGHTS & RECOMMENDATIONS

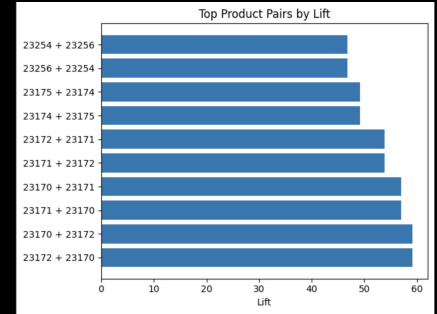
Several strong product pairings were identified.

For example, when item 20711 is purchased, item 20712 is also purchased 54.5% of the time, with a lift of 14.48.

This indicates a highly significant relationship, customers are over 14 times more likely to buy 20712 if they already purchased 20711.

Recommendations:

- Bundle high-lift item pairs (e.g. 20711 + 20712) as promotional combos.
- Display these related products during checkout or in marketing emails to encourage add-on purchases.
- Prioritize cross-sell strategies around top-performing associations to increase average basket size.

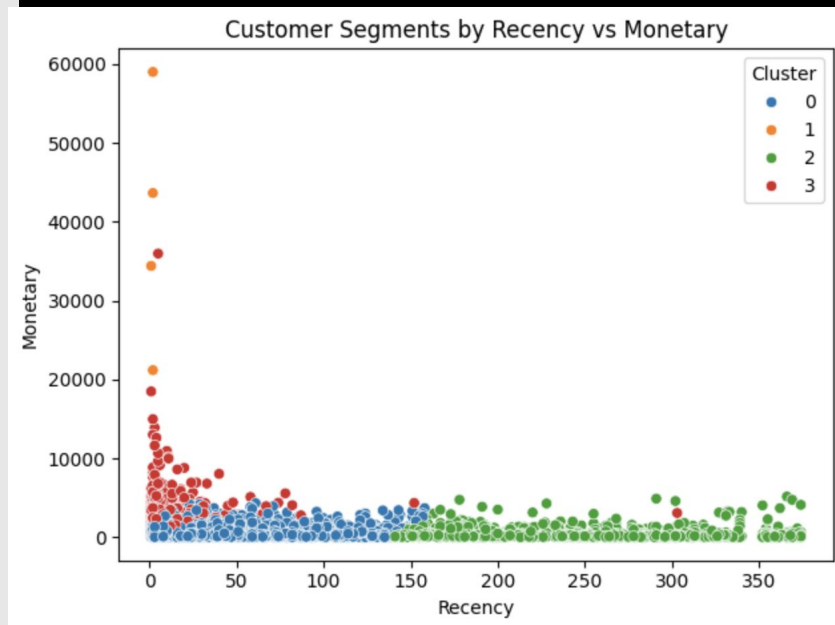


CUSTOMER SEGMENTATION (RFM + K-MEANS)

This task focused on segmenting customers by their purchasing behavior using RFM (Recency, Frequency, Monetary) metrics.

After computing each customer's RFM values, the features were standardized and clustered using K-Means (k=4).

The goal was to uncover behavior-based customer groups to support targeted retention and marketing strategies.



Cluster	Recency	Frequency	Monetary
0	44.7	4.1	702.6
1	1.8	188.5	39551.2
2	247.2	1.7	376.0
3	13.3	22.2	3742.8

SEGMENTATION: INSIGHTS

FOUR DISTINCT CLUSTERS WERE IDENTIFIED:

Cluster 1

Recent, high-frequency, high-spending buyers

Loyal and valuable

Cluster 0

Low frequency and low recency

Low-value customers.

Cluster 3

Moderately active, good spend

Opportunity for upselling

Cluster 2

Inactive, low spend

Likely churned.

Cluster	Recency	Frequency	Monetary
0	44.7	4.1	702.6
1	1.8	188.5	39551.2
2	247.2	1.7	376.0
3	13.3	22.2	3742.8

	Recency	Frequency	Monetary
Cluster			
0	44.7	4.1	702.6
1	1.8	188.5	39551.2
2	247.2	1.7	376.0
3	13.3	22.2	3742.8

Cluster 1

Reward Cluster 1 with exclusive offers to retain loyalty.

Cluster 0

Avoid over-investing in Cluster 0 without strong signals of potential.


SEGMENTATION: RECOMMENDATIONS

Cluster 3

Target Cluster 3 with personalized upsell campaigns.

Cluster 2

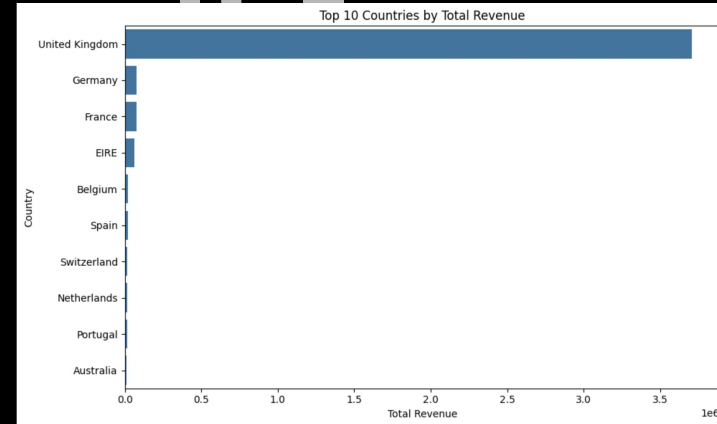
Test reactivation offers on Cluster 2 or deprioritize.



To explore regional performance, transaction data was grouped by country and aggregated by total revenue, unique customers, and transaction volume.

K-Means clustering ($k=3$) was used to group countries with similar purchase behavior to guide international strategy.

COUNTRY-LEVEL SALES CLUSTERING (K-MEANS)



COUNTRY ANALYSIS: INSIGHTS & RECOMMENDATIONS

INTERPRETATION AND RECOMMENDATIONS

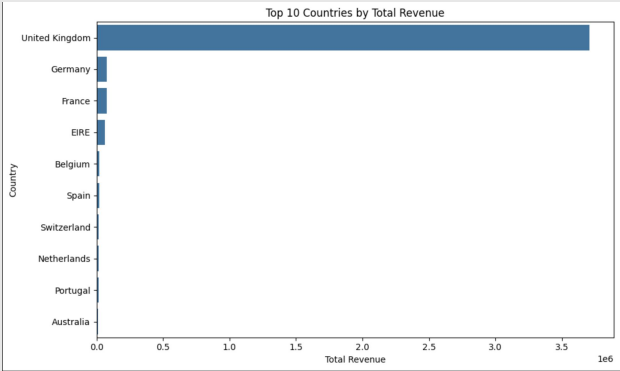


INSIGHTS_

The United Kingdom dominates sales with over \$3.7 million in revenue.

Germany, France, and Ireland show strong mid-tier performance.

Remaining countries have limited activity and contribution.



CLUSTERING REVEALED_

Cluster 1: High-performing (e.g. UK)

Cluster 2: Mid-performing (e.g. Germany, France)

Cluster 0: Low-performing regions

	Country	TotalRevenue
0	United Kingdom	3704989.59
1	Germany	78864.67
2	France	75927.33
3	EIRE	61082.72
4	Belgium	18854.47
5	Spain	18539.21
6	Switzerland	16977.69
7	Netherlands	16835.35
8	Portugal	13530.34
9	Australia	12905.39

RECOMMENDATIONS_

Prioritize continued investment in the UK through promotions and retention.

Explore growth strategies in Germany and France.

Consider lightweight A/B marketing tests in low-activity regions or deprioritize them.

SUMMARY OF FINDINGS

This project applied three machine learning techniques to retail transaction data to generate actionable business insights.

1. APRIORI

Identified high-lift product pairings to drive cross-selling.

2. RFM + K-MEANS

Segmented customers to enable personalized retention and upsell strategies.

3. COUNTRY CLUSTERING

Uncovered high-, mid-, and low-performing regions for focused expansion or testing.

Together, these models support data-driven decisions that align with marketing, product bundling, and regional investment strategies.

THANK YOU!

