NAME : **GANESH S**

REGISTER NO : **727723EUCS059**

EMAIL ID : **727723eucs059@skcet.ac.in**

PROJECT NAME : **Customer Segmentation & Market Basket Intelligence Platform**

GITHUB LINK : **https://github.com/GaneshLathin/Customer-Segmentation-Market-Basket-Intelligence-Platform**

Customer Segmentation & Market Basket Intelligence Platform

Problem Statement

An e-commerce business generates hundreds of thousands of transaction records annually across thousands of customers and products. However, without structured analysis, it becomes extremely difficult to identify high-value customers, detect at-risk customers, understand purchasing patterns, or create targeted marketing strategies.

Traditional rule-based segmentation techniques fail to capture complex behavioral patterns hidden in large transactional datasets. Manual analysis is inefficient and prone to oversight, especially when dealing with multi-dimensional customer features such as recency, frequency, and monetary value.

Therefore, there is a need for an intelligent analytics platform that applies unsupervised machine learning algorithms to automatically discover meaningful customer segments, reduce data complexity, uncover product associations, and generate actionable marketing personas. This project addresses that need through the development of SegmentIQ.

Objectives

The main objectives of this project are:

1. To ingest and clean raw transactional data into a structured customer-level dataset.
2. To engineer RFM (Recency, Frequency, Monetary) features along with additional behavioral metrics.
3. To apply multiple unsupervised clustering algorithms for automated customer segmentation.

4. To reduce high-dimensional customer data using dimensionality reduction techniques.
5. To perform market basket analysis to identify product association patterns.
6. To generate interpretable marketing personas based on cluster characteristics.
7. To build an interactive full-stack web dashboard for real-time analytics visualization.

Model Implementation :

Unsupervised Machine Learning for Customer Segmentation:

SegmentIQ applies multiple clustering and analytical techniques:

1. **K-Means Clustering:**
   Uses centroid-based partitioning to divide customers into optimal segments based on behavioral similarity.
2. **Agglomerative Hierarchical Clustering:**
   Builds a bottom-up dendrogram structure to represent hierarchical merging of customers.
3. **DBSCAN (Density-Based Spatial Clustering):**
   Identifies arbitrarily shaped clusters and detects anomalous customers as noise points.

Steps involved:

- Perform exploratory data analysis (EDA)
- Aggregate invoice-level data into customer-level features
- Engineer RFM and derived behavioral metrics
- Apply log transformation (log1p) to reduce skewness
- Standardize features using z-score normalization
- Determine optimal cluster parameters using silhouette score
- Train clustering models
- Interpret cluster characteristics

Feature Engineering:

Customer-level features include:

- Recency (Days since last purchase)
- Frequency (Number of unique invoices)
- Monetary Value (Total spending)
- Average Order Value
- Total Items Purchased
- Unique Products Bought
- Average Basket Size

Dimensionality Reduction:

To visualize high-dimensional customer data:

1. **PCA (Principal Component Analysis):**
   Reduces multi-dimensional feature space into principal components while retaining maximum variance.
2. **LDA (Linear Discriminant Analysis):**
   Uses cluster labels as targets to maximize separation between discovered segments.

Techniques used:

- Explained variance ratio analysis
- Component loading interpretation
- 2D and 3D visualization of clusters

Market Basket Analysis:

The Apriori algorithm is applied to extract frequent itemsets and association rules from transaction data.

• Identify frequently co-purchased products

• Generate rules using support, confidence, and lift

• Rank product relationships by lift ratio

• Visualize product co-occurrence using heatmaps

Hyperparameter Tuning:

The following parameters are optimized:

- k (Number of clusters in K-Means)
- eps (Neighborhood radius in DBSCAN)
- min_samples (Minimum points for density cluster)
- min_support (Apriori support threshold)

Techniques used:

- Elbow method for inertia analysis
- Silhouette score comparison
- Dendrogram analysis
- Lift-based ranking of association rules

Performance Evaluation:

Clustering models are evaluated using unsupervised performance metrics.

The following metrics are used:

- Silhouette Score
- Inertia (Within-cluster sum of squares)
- Davies-Bouldin Index
- Explained Variance Ratio (PCA)
- Support, Confidence, and Lift (Apriori)

System Architecture:

Backend:

- Python 3.11
- FastAPI framework
- scikit-learn, SciPy, mlxtend
- Automatic dataset download from UCI repository
- REST API endpoints for clustering and analytics

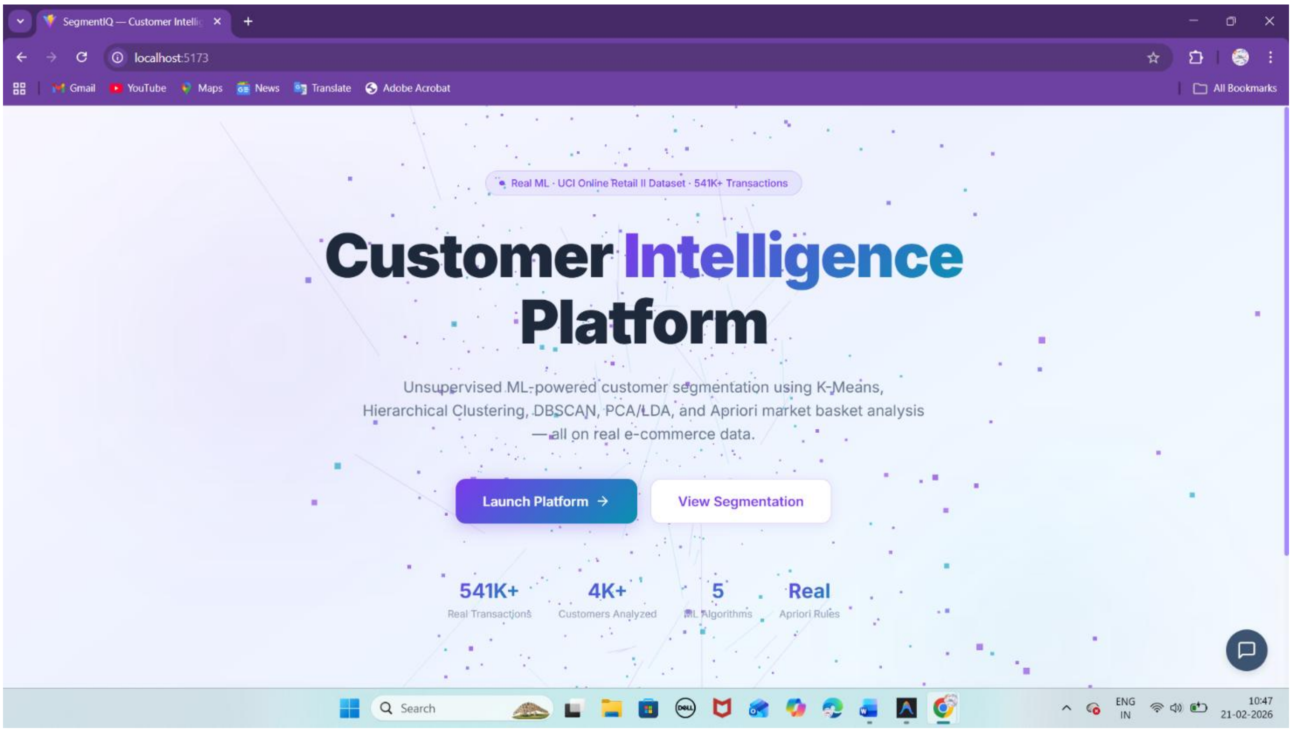
Frontend:

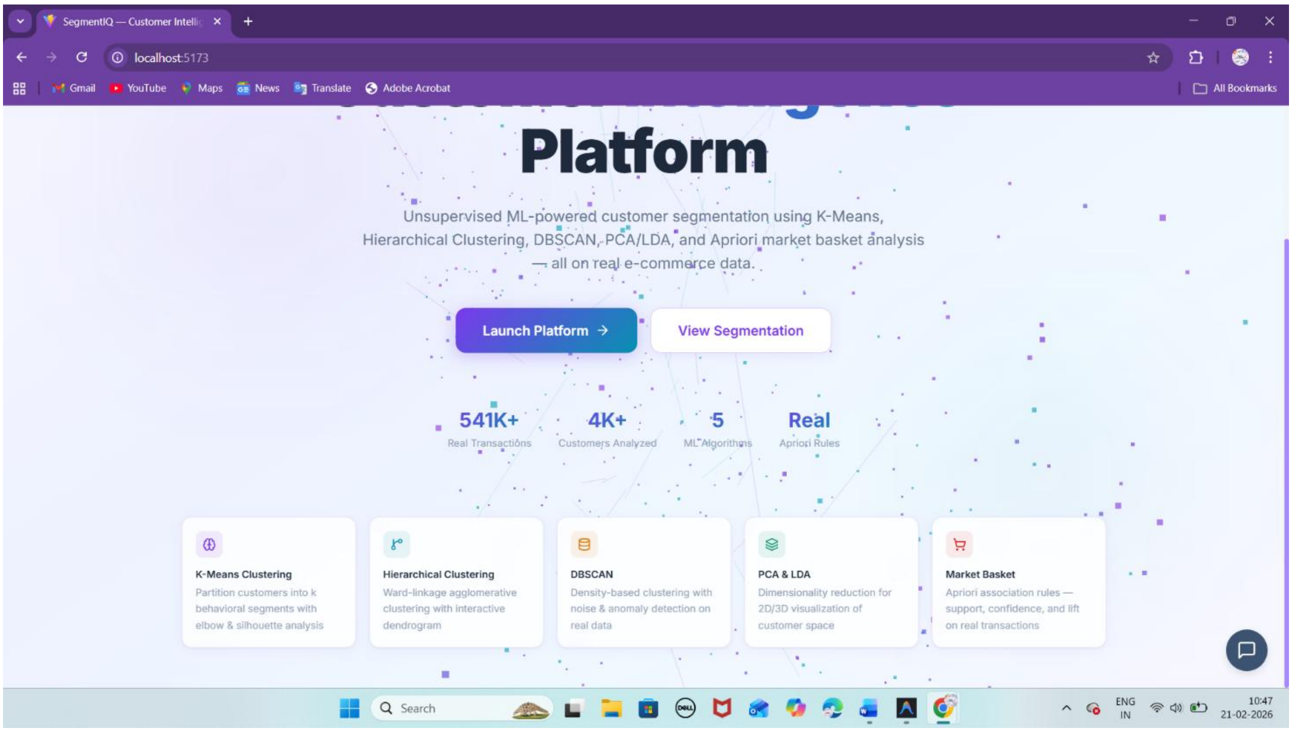
- React 18 with Vite
- Tailwind CSS styling
- Recharts for data visualization
- Three.js for 3D background
- Framer Motion & GSAP for animations
- Axios for API communication

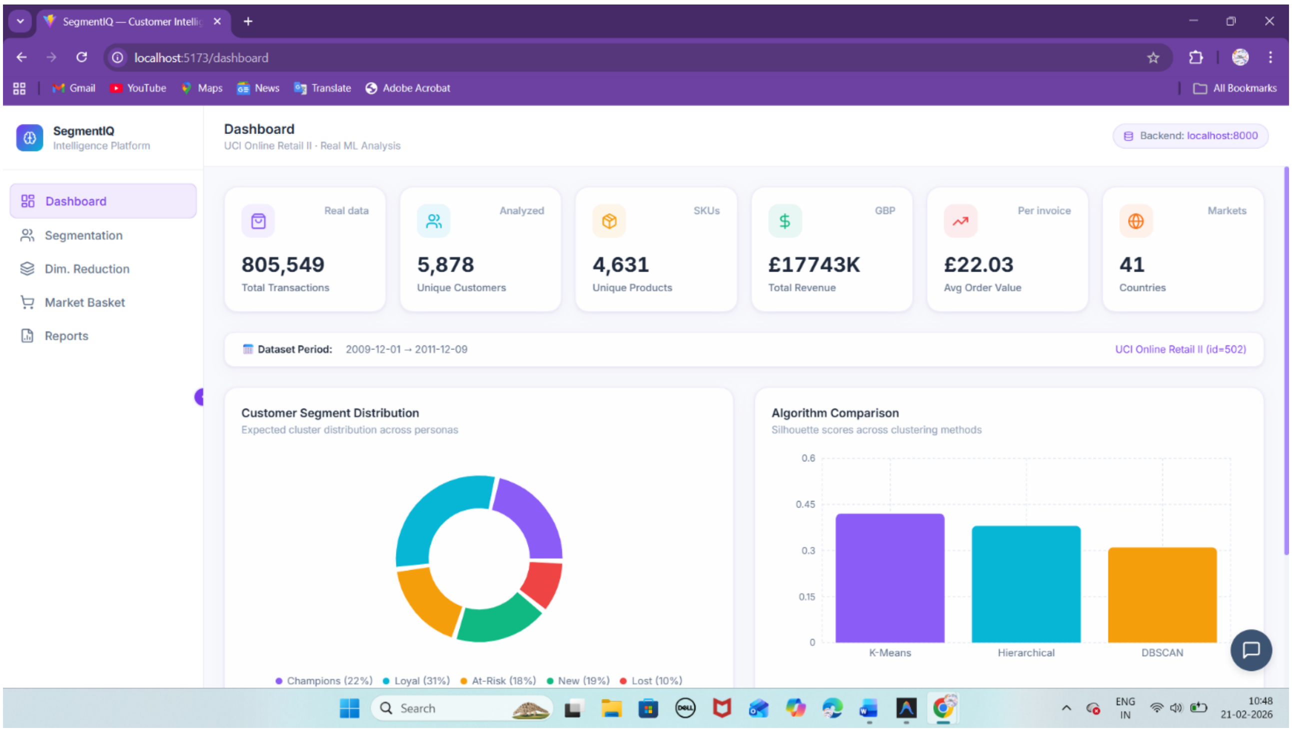
Results and Outcomes:

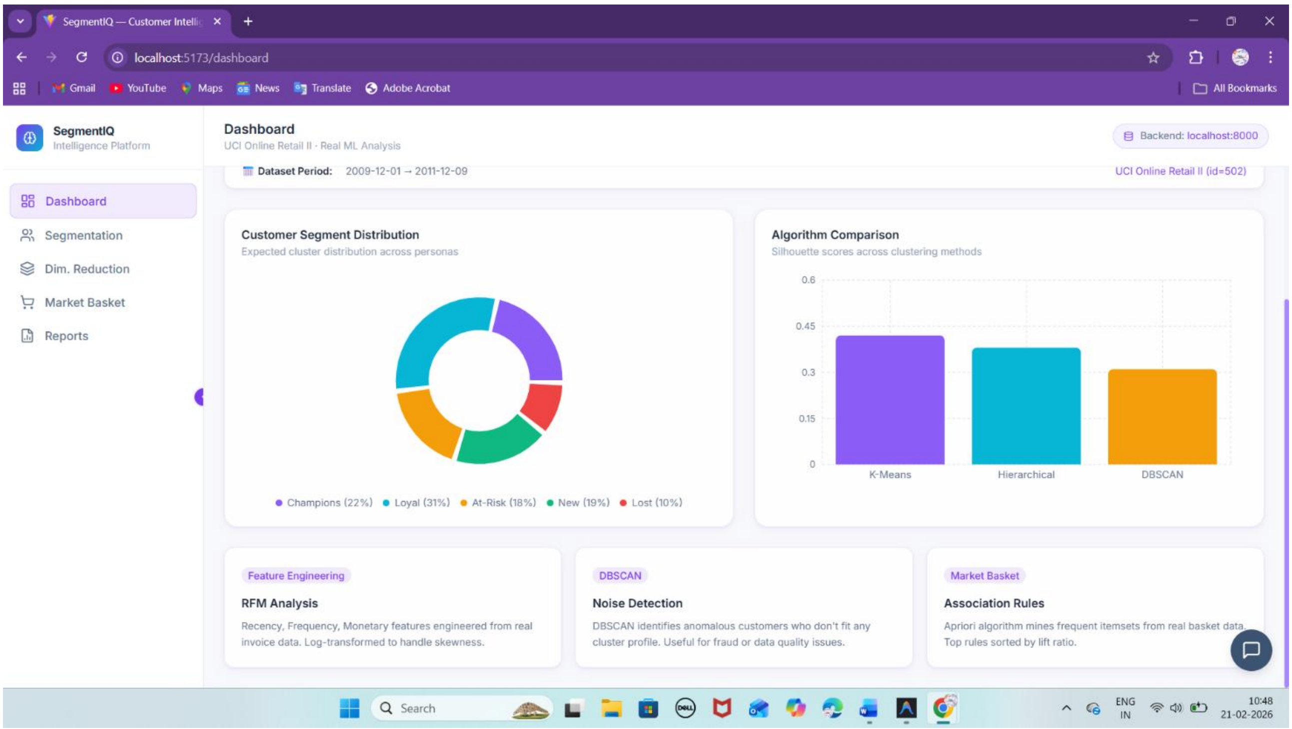
- Successful segmentation of customers into meaningful behavioral groups
- Silhouette score of approximately 0.42 using K-Means
- Identification of high-value "Champion" customers
- Detection of 5–12% anomalous customers using DBSCAN
- PCA explaining approximately 65–70% variance in first two components
- Discovery of high-lift product association rules
- Generation of actionable marketing personas with campaign recommendations
- Interactive real-time dashboard for business decision support

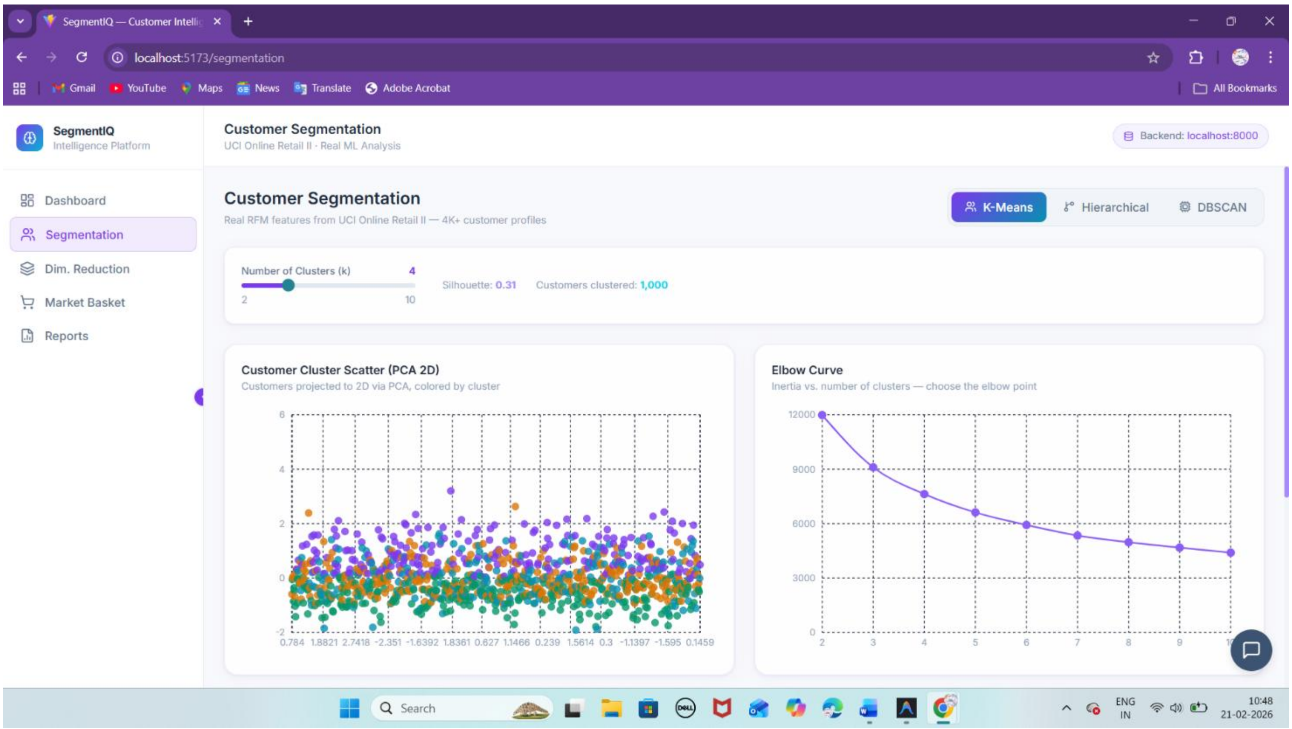
Output:

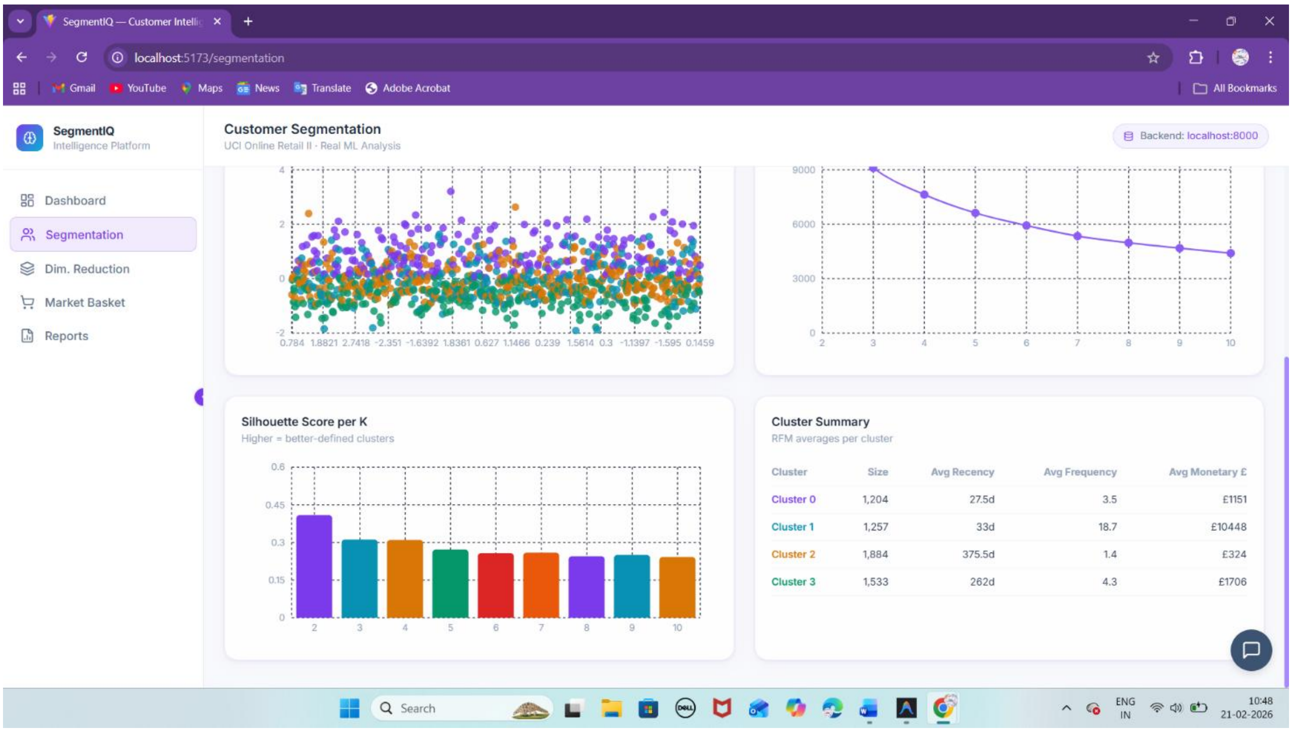

SegmentIQ — Customer Intelli
localhost:5173
Gmail
YouTube
Maps
News
Translate
Adobe Acrobat
All Bookmarks
Real ML · UCI Online Retail II Dataset · 541K+ Transactions
Customer Intelligence Platform
Unsupervised ML-powered customer segmentation using K-Means, Hierarchical Clustering, DBSCAN, PCA/LDA, and Apriori market basket analysis — all on real e-commerce data.
Launch Platform
View Segmentation
541K+
Real Transactions
4K+
Customers Analyzed
5
ML Algorithms
Real
Apriori Rules
Search
ENG
IN
10:47
21-02-2026

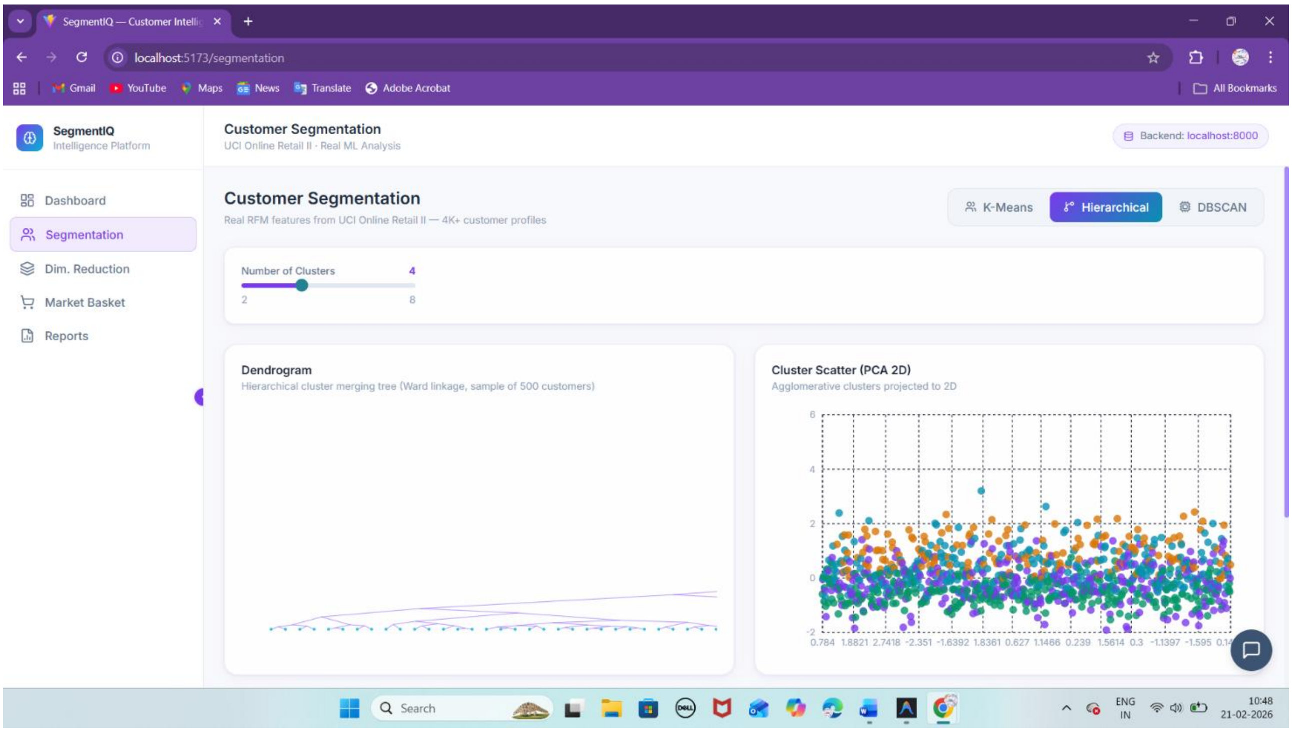

SegmentIQ — Customer Intelli
localhost:5173
Gmail
YouTube
Maps
News
Translate
Adobe Acrobat
All Bookmarks
Platform
Unsupervised ML-powered customer segmentation using K-Means, Hierarchical Clustering, DBSCAN, PCA/LDA, and Apriori market basket analysis — all on real e-commerce data.
Launch Platform
View Segmentation
541K+
Real Transactions
4K+
Customers Analyzed
5
ML Algorithms
Real
Apriori Rules
K-Means Clustering
Partition customers into k behavioral segments with elbow & silhouette analysis
Hierarchical Clustering
Ward-linkage agglomerative clustering with interactive dendrogram
DBSCAN
Density-based clustering with noise & anomaly detection on real data
PCA & LDA
Dimensionality reduction for 2D/3D visualization of customer space
Market Basket
Apriori association rules — support, confidence, and lift on real transactions
Search
ENG
IN
10:47
21-02-2026


SegmentIQ — Customer Intelli
localhost:5173/dashboard
Gmail
YouTube
Maps
News
Translate
Adobe Acrobat
All Bookmarks
SegmentIQ
Intelligence Platform
Dashboard
Segmentation
Dim. Reduction
Market Basket
Reports
Dashboard
UCI Online Retail II · Real ML Analysis
Backend: localhost:8000
Real data
805,549
Total Transactions
Analyzed
5,878
Unique Customers
SKUs
4,631
Unique Products
GBP
£17743K
Total Revenue
Per invoice
£22.03
Avg Order Value
Markets
41
Countries
Dataset Period: 2009-12-01 → 2011-12-09
UCI Online Retail II (id=502)
Customer Segment Distribution
Expected cluster distribution across personas
Champions (22%)
Loyal (31%)
At-Risk (18%)
New (19%)
Lost (10%)
Algorithm Comparison
Silhouette scores across clustering methods
0.6
0.45
0.3
0.15
0
K-Means
Hierarchical
DBSCAN
Search
ENG
IN
10:48
21-02-2026

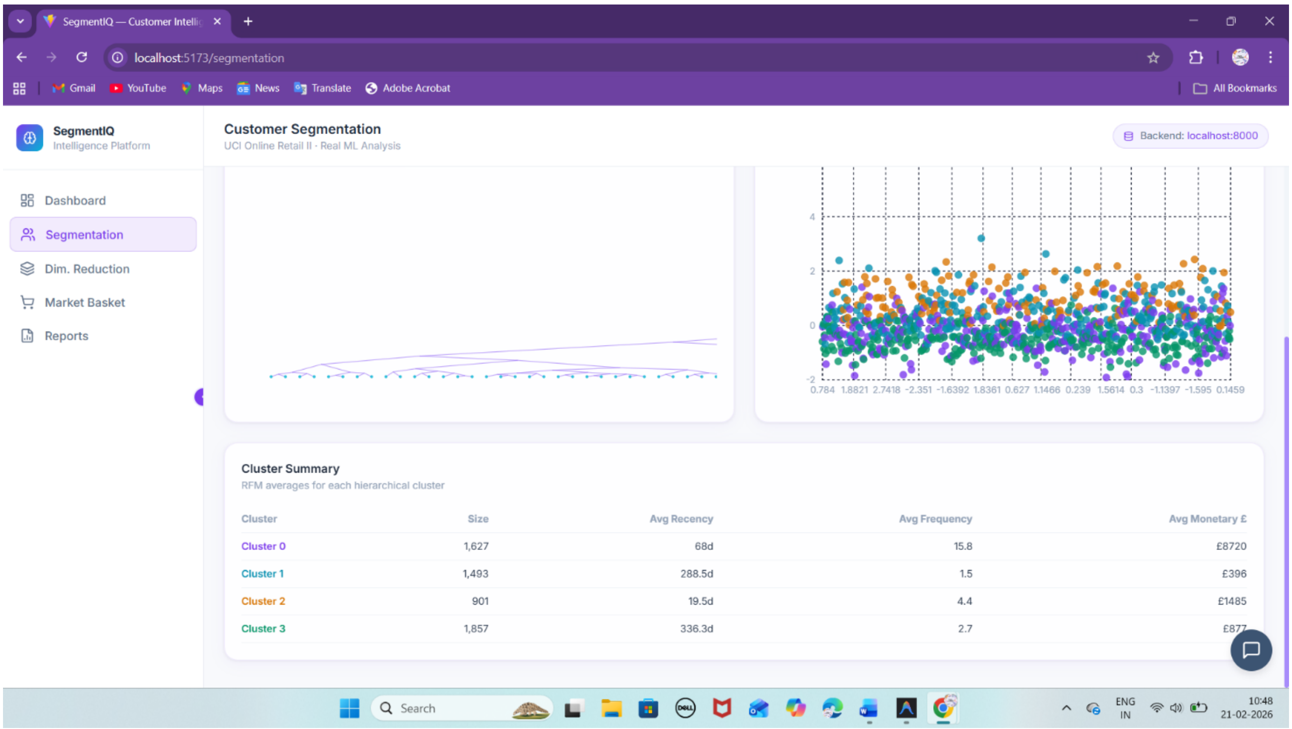

SegmentIQ — Customer Intelli
localhost:5173/dashboard
Gmail
YouTube
Maps
News
Translate
Adobe Acrobat
All Bookmarks
SegmentIQ
Intelligence Platform
Dashboard
Segmentation
Dim. Reduction
Market Basket
Reports
Dashboard
UCI Online Retail II · Real ML Analysis
Backend: localhost:8000
Dataset Period: 2009-12-01 → 2011-12-09
UCI Online Retail II (id=502)
Customer Segment Distribution
Expected cluster distribution across personas
Champions (22%)
Loyal (31%)
At-Risk (18%)
New (19%)
Lost (10%)
Algorithm Comparison
Silhouette scores across clustering methods
0.6
0.45
0.3
0.15
0
K-Means
Hierarchical
DBSCAN
Feature Engineering
RFM Analysis
Recency, Frequency, Monetary features engineered from real invoice data. Log-transformed to handle skewness.
DBSCAN
Noise Detection
DBSCAN identifies anomalous customers who don't fit any cluster profile. Useful for fraud or data quality issues.
Market Basket
Association Rules
Apriori algorithm mines frequent itemsets from real basket data. Top rules sorted by lift ratio.
Search
ENG
IN
10:48
21-02-2026


SegmentIQ — Customer Intelli
localhost:5173/segmentation
Gmail YouTube Maps News Translate Adobe Acrobat
All Bookmarks
SegmentIQ
Intelligence Platform
Customer Segmentation
UCI Online Retail II · Real ML Analysis
Backend: localhost:8000
Dashboard
Segmentation
Dim. Reduction
Market Basket
Reports
Customer Segmentation
Real RFM features from UCI Online Retail II — 4K+ customer profiles
K-Means
Hierarchical
DBSCAN
Number of Clusters (k) 4
2 10
Silhouette: 0.31
Customers clustered: 1,000
Customer Cluster Scatter (PCA 2D)
Customers projected to 2D via PCA, colored by cluster
Elbow Curve
Inertia vs. number of clusters — choose the elbow point
Search
ENG IN
10:48 21-02-2026

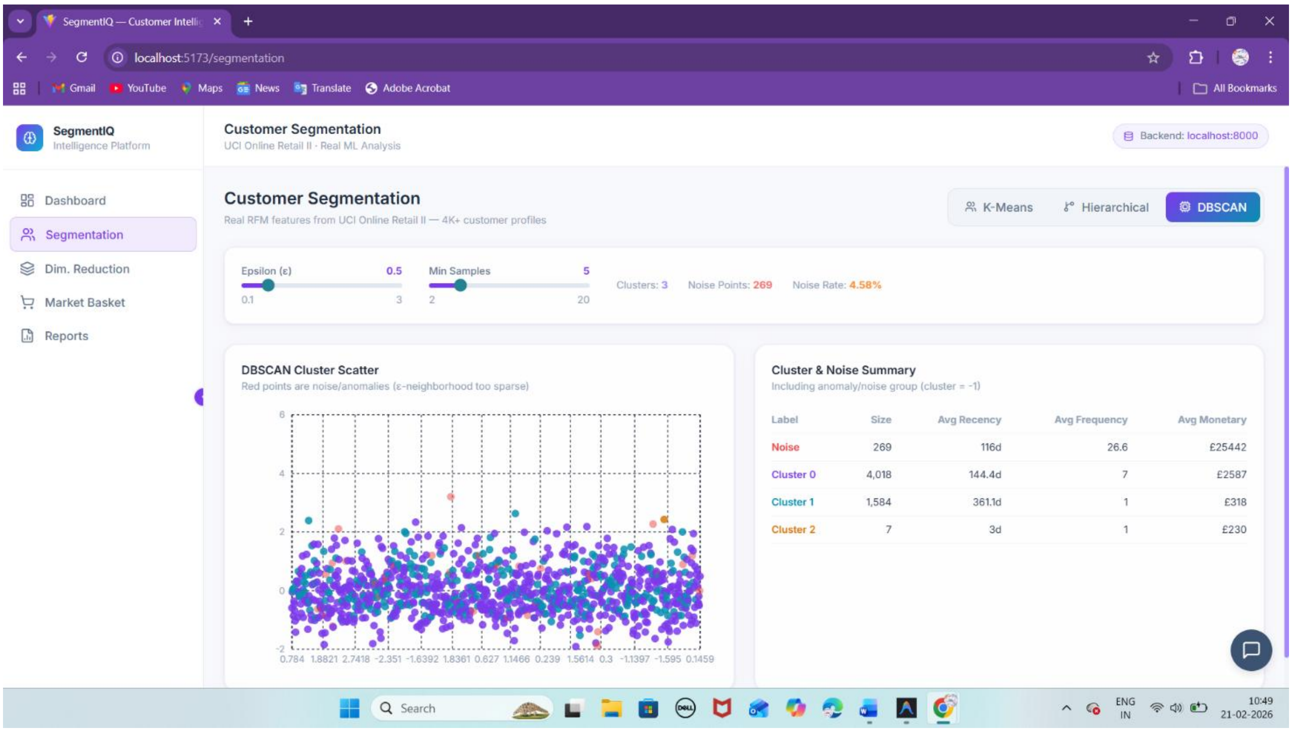

SegmentIQ — Customer Intelli
localhost:5173/segmentation
Customer Segmentation
UCI Online Retail II · Real ML Analysis
Backend: localhost:8000
Dashboard
Segmentation
Dim. Reduction
Market Basket
Reports
Silhouette Score per K
Higher = better-defined clusters
Cluster Summary
RFM averages per cluster
Cluster Size Avg Recency Avg Frequency Avg Monetary £
Cluster 0 1,204 27.5d 3.5 £1151
Cluster 1 1,257 33d 18.7 £10448
Cluster 2 1,884 375.5d 1.4 £324
Cluster 3 1,533 262d 4.3 £1706
Search
ENG IN
10:48 21-02-2026

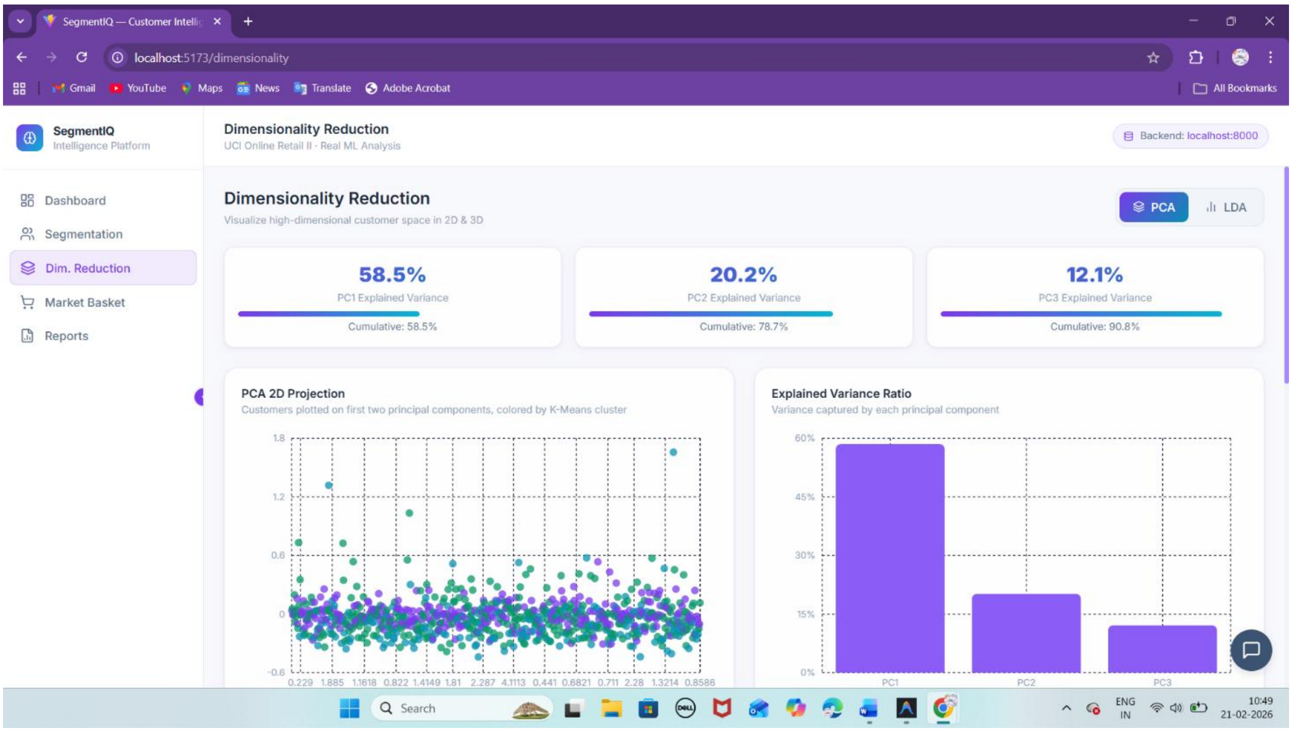


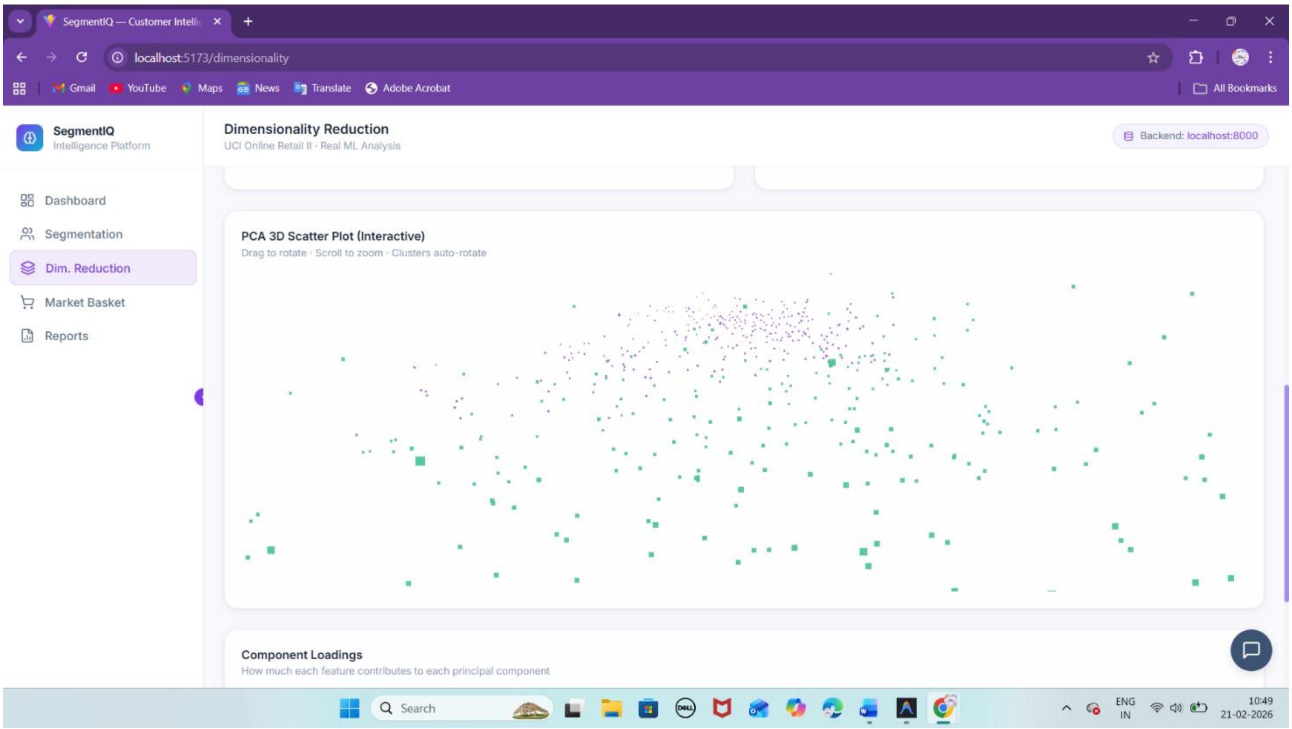
SegmentIQ — Customer Intellig
localhost:5173/segmentation
Gmail
YouTube
Maps
News
Translate
Adobe Acrobat
All Bookmarks
SegmentIQ
Intelligence Platform
Customer Segmentation
UCI Online Retail II · Real ML Analysis
Backend: localhost:8000
Dashboard
Segmentation
Dim. Reduction
Market Basket
Reports
Customer Segmentation
Real RFM features from UCI Online Retail II — 4K+ customer profiles
K-Means
Hierarchical
DBSCAN
Number of Clusters
4
2
8
Dendrogram
Hierarchical cluster merging tree (Ward linkage, sample of 500 customers)
Cluster Scatter (PCA 2D)
Agglomerative clusters projected to 2D
Search
ENG
IN
10:48
21-02-2026

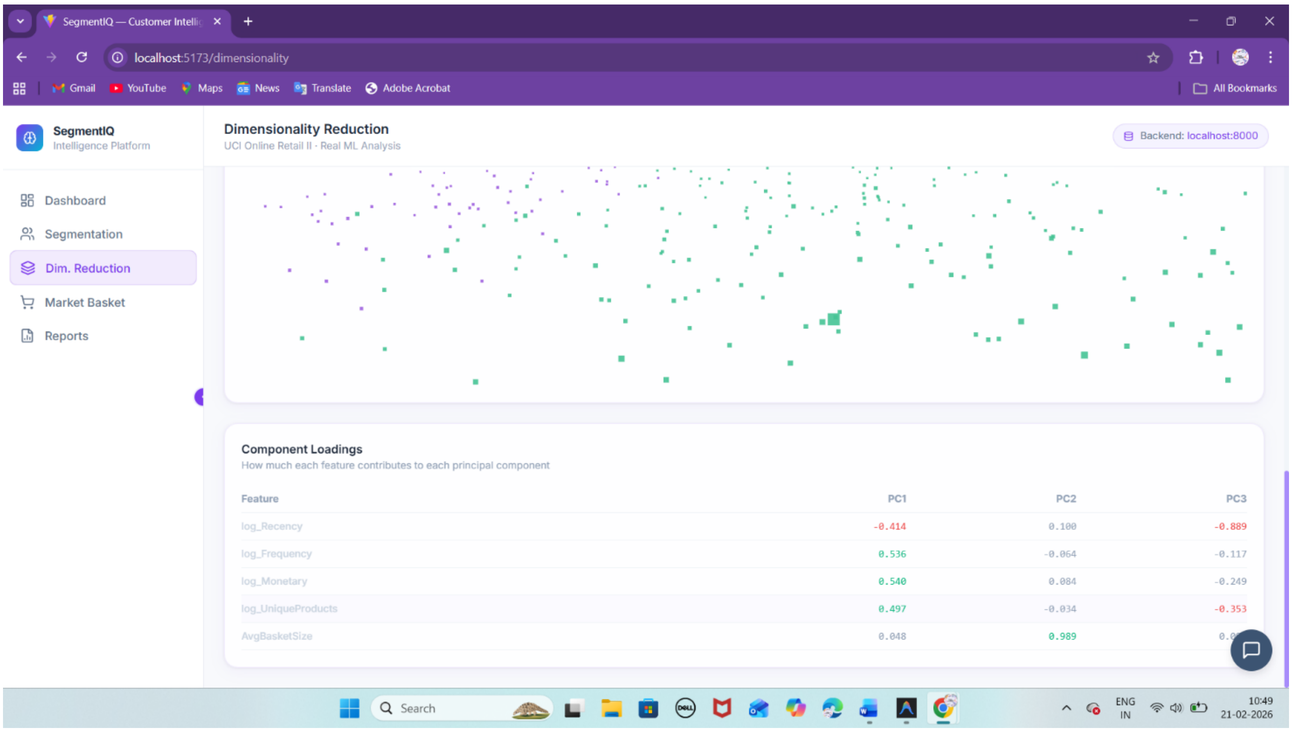


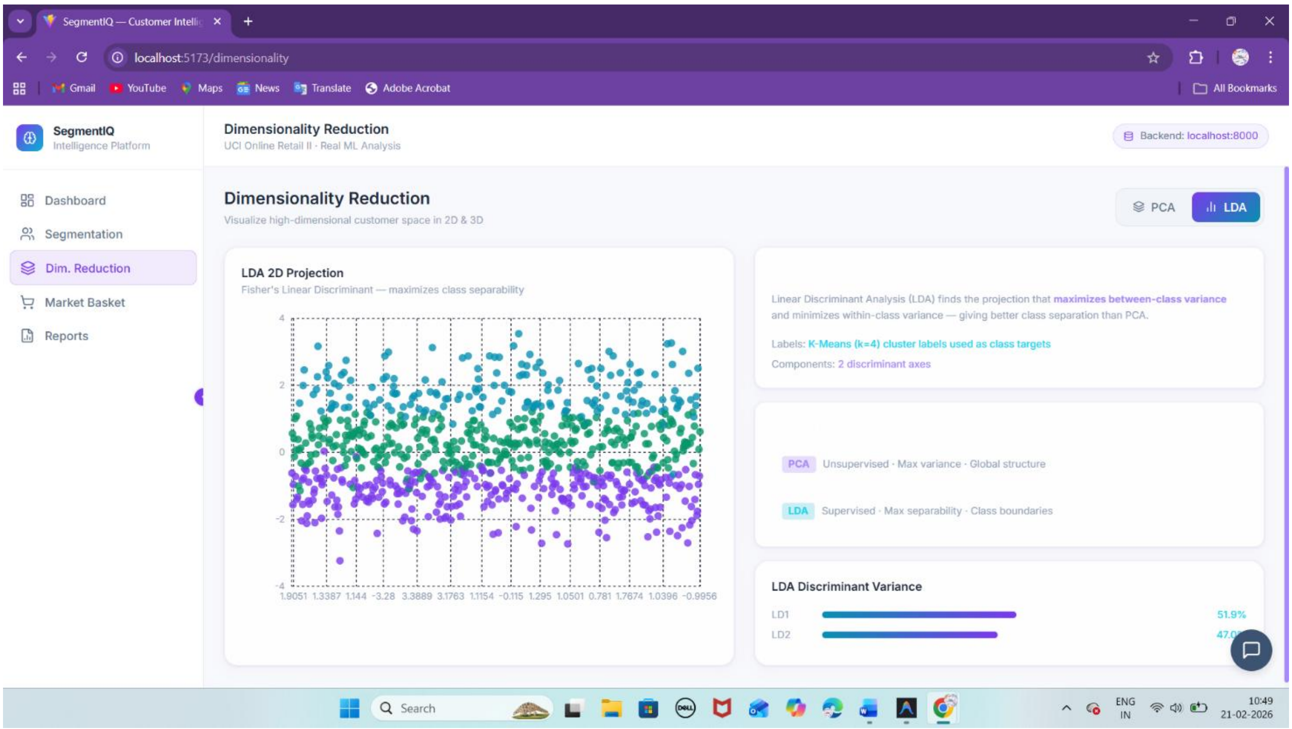
SegmentIQ — Customer Intellig
localhost:5173/segmentation
SegmentIQ
Intelligence Platform
Customer Segmentation
UCI Online Retail II · Real ML Analysis
Backend: localhost:8000
Dashboard
Segmentation
Dim. Reduction
Market Basket
Reports
Cluster Summary
RFM averages for each hierarchical cluster
Cluster | Size | Avg Recency | Avg Frequency | Avg Monetary £
Cluster 0 | 1,627 | 68d | 15.8 | £8720
Cluster 1 | 1,493 | 288.5d | 1.5 | £396
Cluster 2 | 901 | 19.5d | 4.4 | £1485
Cluster 3 | 1,857 | 336.3d | 2.7 | £877
Search
ENG
IN
10:48
21-02-2026


SegmentIQ — Customer Intelli
localhost:5173/segmentation
SegmentIQ
Intelligence Platform
Customer Segmentation
UCI Online Retail II · Real ML Analysis
Backend: localhost:8000
Dashboard
Segmentation
Dim. Reduction
Market Basket
Reports
Customer Segmentation
Real RFM features from UCI Online Retail II — 4K+ customer profiles
K-Means
Hierarchical
DBSCAN
Epsilon (ε) 0.5
Min Samples 5
Clusters: 3
Noise Points: 269
Noise Rate: 4.58%
DBSCAN Cluster Scatter
Red points are noise/anomalies (ε-neighborhood too sparse)
Cluster & Noise Summary
Including anomaly/noise group (cluster = -1)
Label | Size | Avg Recency | Avg Frequency | Avg Monetary
Noise | 269 | 116d | 26.6 | £25442
Cluster 0 | 4,018 | 144.4d | 7 | £2587
Cluster 1 | 1,584 | 361.1d | 1 | £318
Cluster 2 | 7 | 3d | 1 | £230

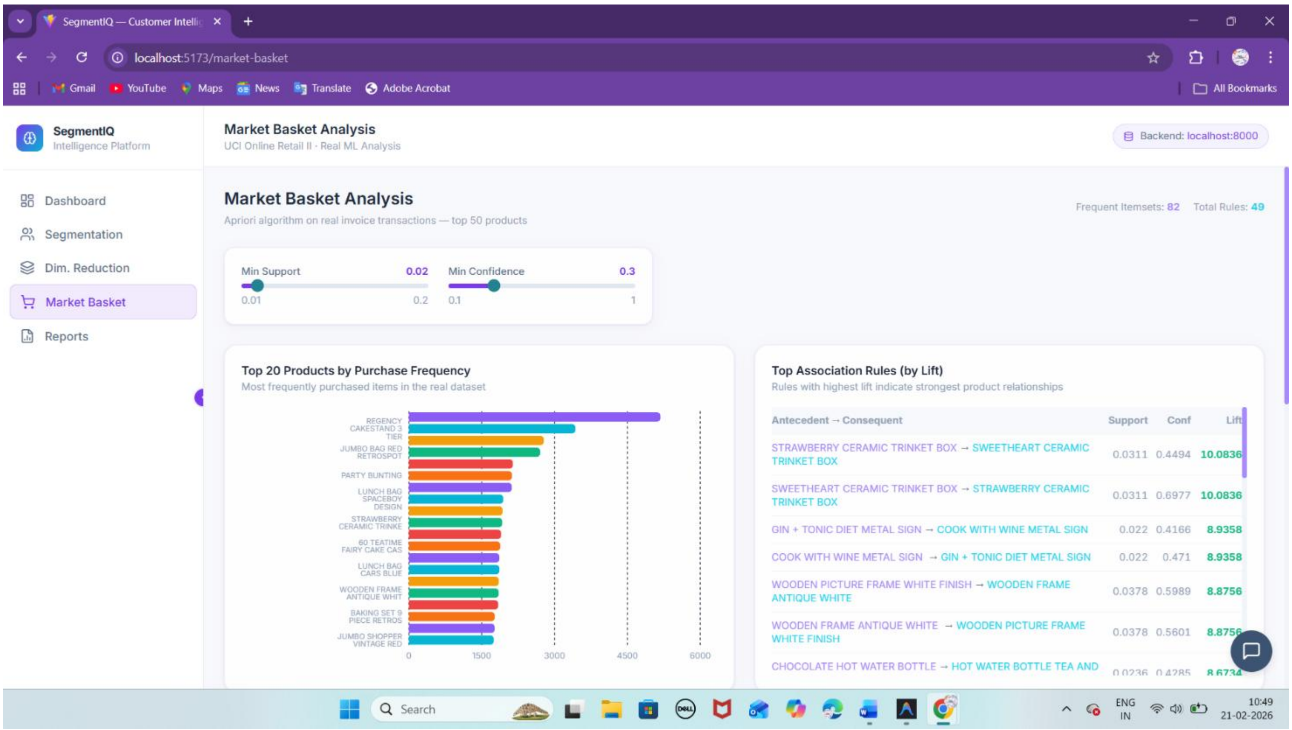

SegmentIQ — Customer Intelli
localhost:5173/dimensionality
SegmentIQ
Intelligence Platform
Dimensionality Reduction
UCI Online Retail II · Real ML Analysis
Backend: localhost:8000
Dashboard
Segmentation
Dim. Reduction
Market Basket
Reports
Dimensionality Reduction
Visualize high-dimensional customer space in 2D & 3D
PCA
LDA
58.5%
PC1 Explained Variance
Cumulative: 58.5%
20.2%
PC2 Explained Variance
Cumulative: 78.7%
12.1%
PC3 Explained Variance
Cumulative: 90.8%
PCA 2D Projection
Customers plotted on first two principal components, colored by K-Means cluster
Explained Variance Ratio
Variance captured by each principal component
PC1
PC2
PC3


SegmentIQ — Customer Intelli
localhost:5173/dimensionality
Gmail
YouTube
Maps
News
Translate
Adobe Acrobat
All Bookmarks
SegmentIQ
Intelligence Platform
Dashboard
Segmentation
Dim. Reduction
Market Basket
Reports
Dimensionality Reduction
UCI Online Retail II · Real ML Analysis
Backend: localhost:8000
PCA 3D Scatter Plot (Interactive)
Drag to rotate · Scroll to zoom · Clusters auto-rotate
Component Loadings
How much each feature contributes to each principal component
Search
ENG
IN
10:49
21-02-2026

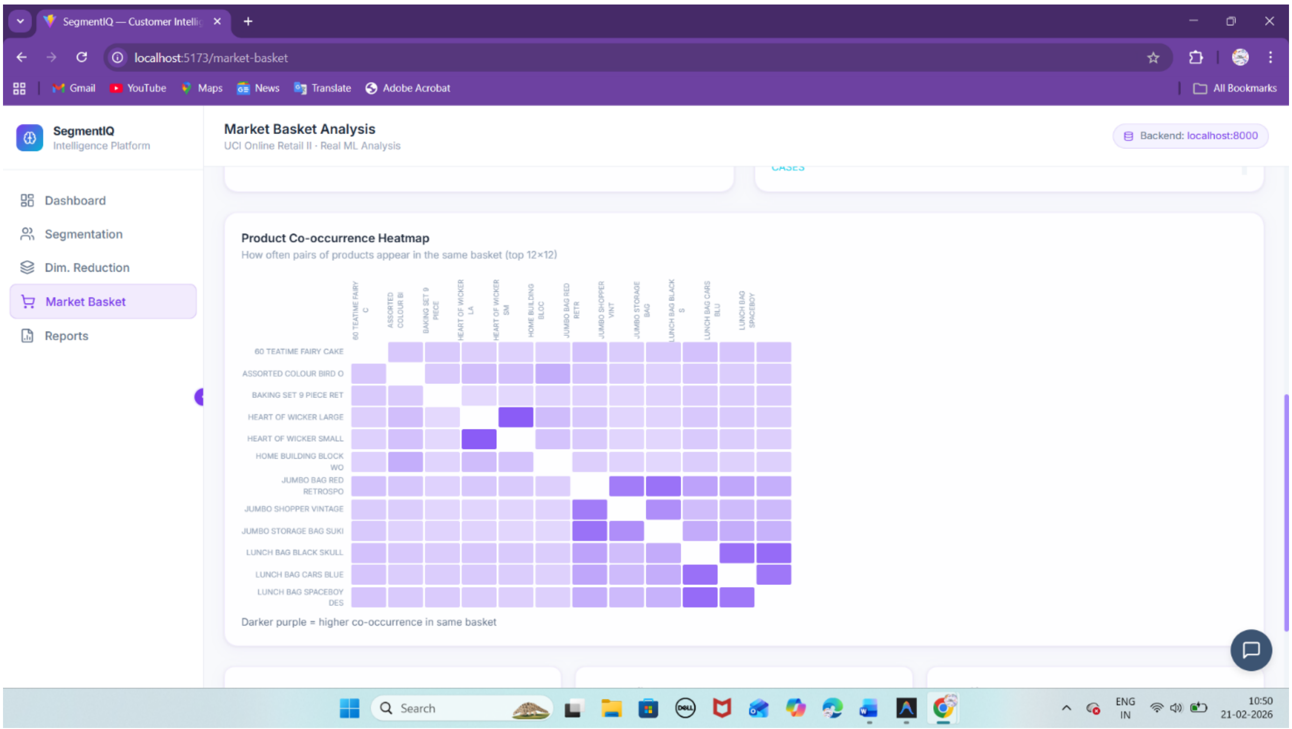

SegmentIQ — Customer Intelli
localhost:5173/dimensionality
Gmail
YouTube
Maps
News
Translate
Adobe Acrobat
All Bookmarks
SegmentIQ
Intelligence Platform
Dashboard
Segmentation
Dim. Reduction
Market Basket
Reports
Dimensionality Reduction
UCI Online Retail II · Real ML Analysis
Backend: localhost:8000
Component Loadings
How much each feature contributes to each principal component
Feature PC1 PC2 PC3
log_Recency -0.414 0.100 -0.889
log_Frequency 0.536 -0.064 -0.117
log_Monetary 0.540 0.084 -0.249
log_UniqueProducts 0.497 -0.034 -0.353
AvgBasketSize 0.048 0.989
Search
ENG
IN
10:49
21-02-2026

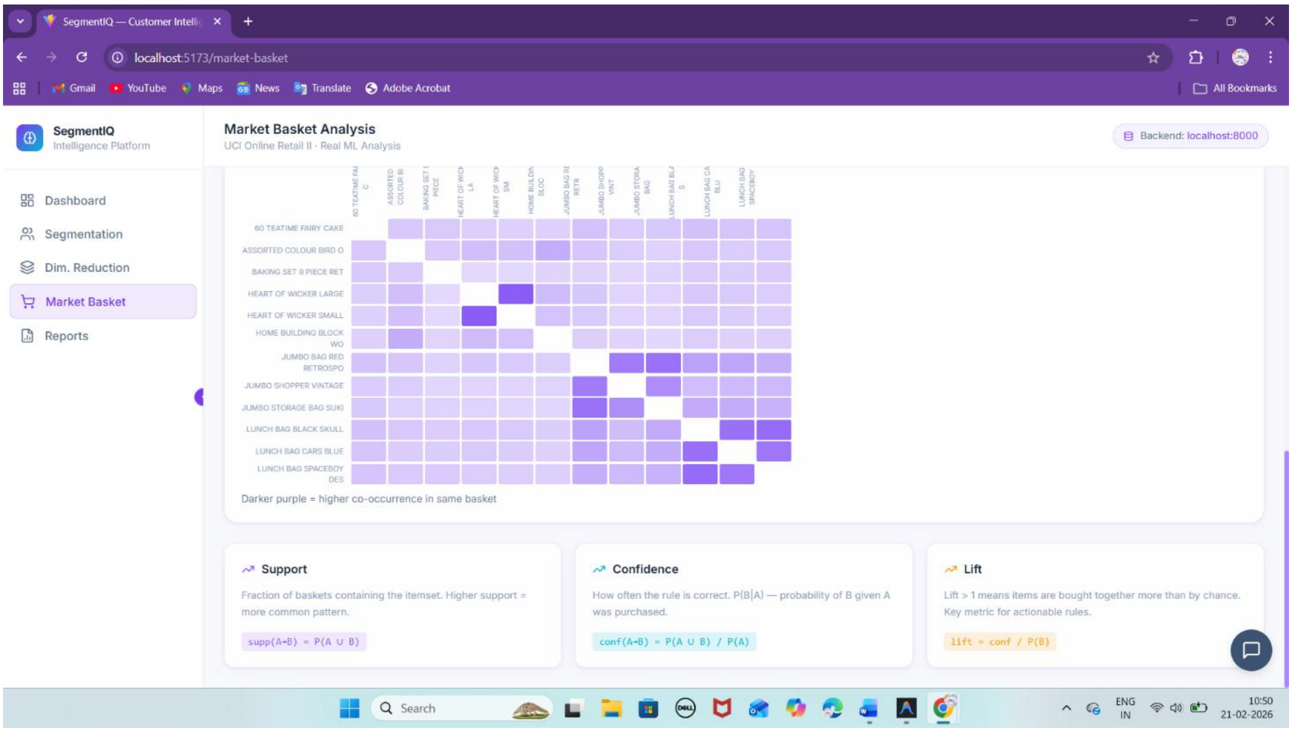


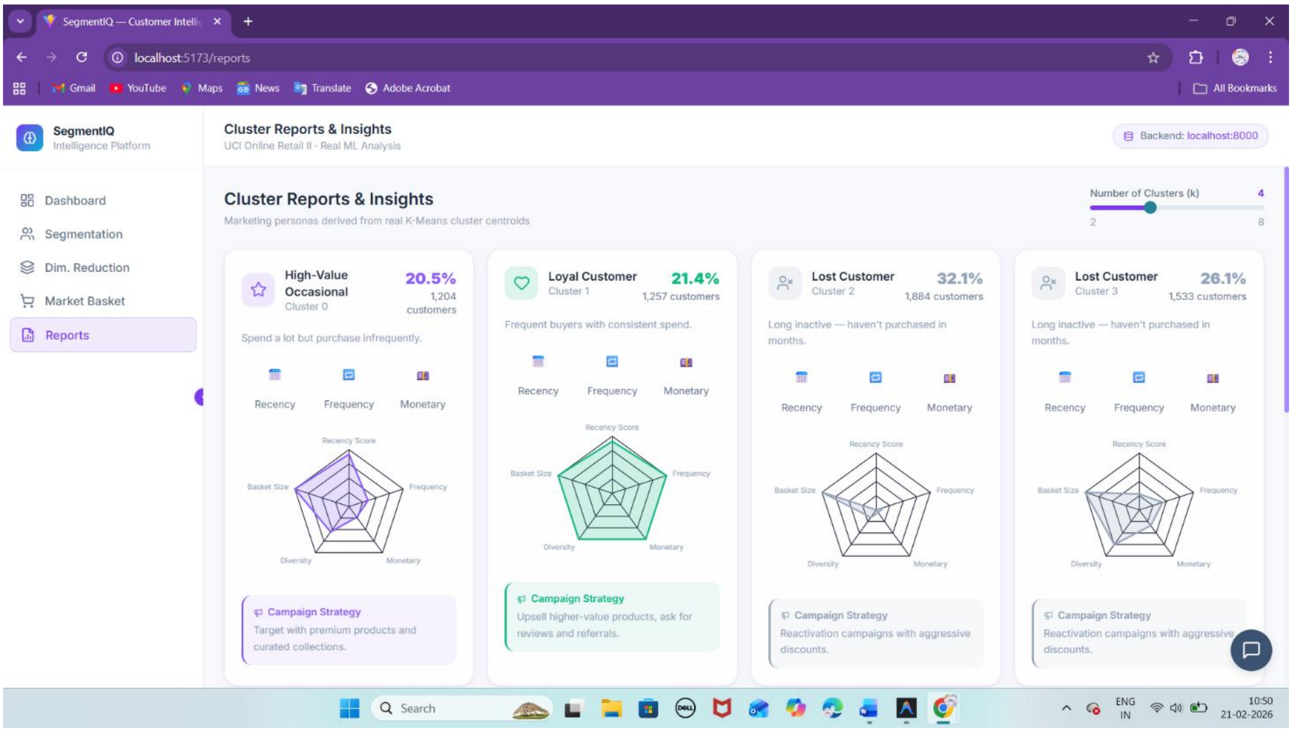
Dimensionality Reduction
UCI Online Retail II · Real ML Analysis
Visualize high-dimensional customer space in 2D & 3D
LDA 2D Projection
Fisher's Linear Discriminant — maximizes class separability
Linear Discriminant Analysis (LDA) finds the projection that maximizes between-class variance and minimizes within-class variance — giving better class separation than PCA.
Labels: K-Means (k=4) cluster labels used as class targets
Components: 2 discriminant axes
PCA Unsupervised · Max variance · Global structure
LDA Supervised · Max separability · Class boundaries
LDA Discriminant Variance
LD1 51.9%

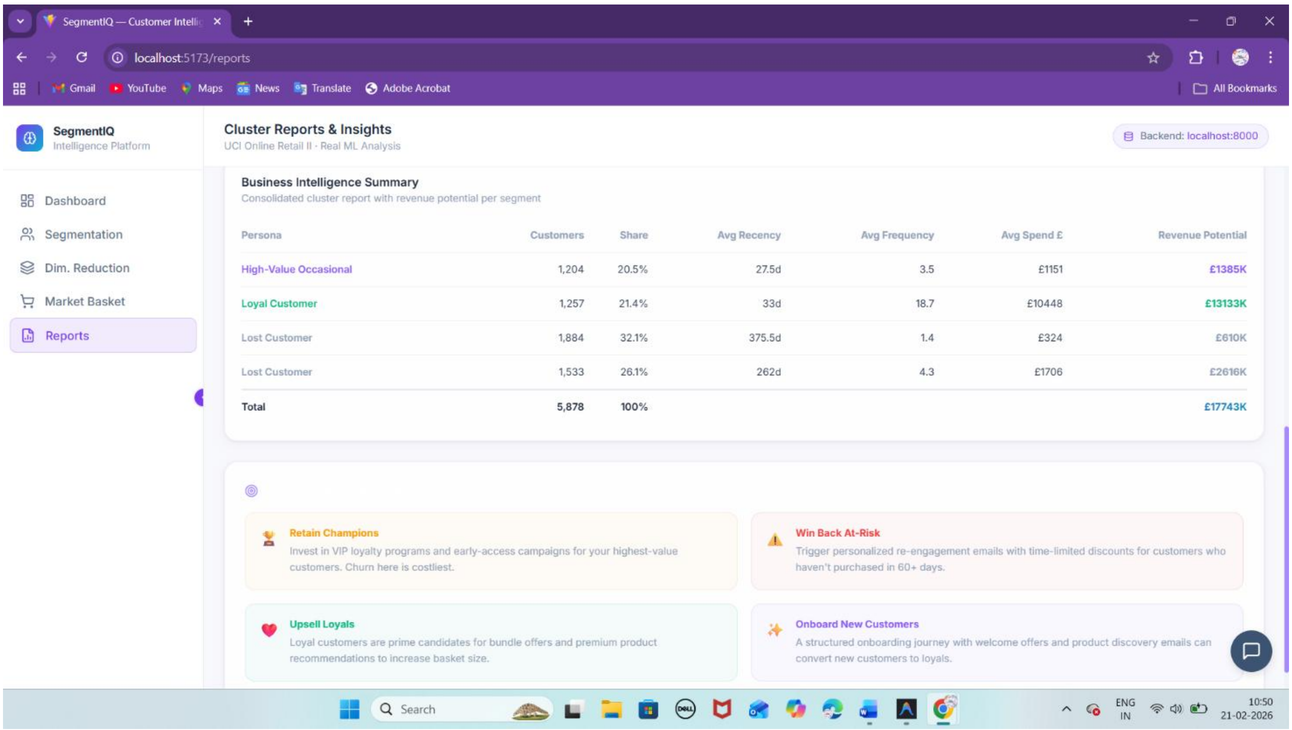


Market Basket Analysis
Apriori algorithm on real invoice transactions — top 50 products
Frequent Itemsets: 82 Total Rules: 49
Min Support 0.02
Min Confidence 0.3
Top 20 Products by Purchase Frequency
Most frequently purchased items in the real dataset
Top Association Rules (by Lift)
Rules with highest lift indicate strongest product relationships
STRAWBERRY CERAMIC TRINKET BOX → SWEETHEART CERAMIC TRINKET BOX 0.0311 0.4494 10.0836
SWEETHEART CERAMIC TRINKET BOX → STRAWBERRY CERAMIC TRINKET BOX 0.0311 0.6977 10.0836
GIN + TONIC DIET METAL SIGN → COOK WITH WINE METAL SIGN 0.022 0.4166 8.9358
COOK WITH WINE METAL SIGN → GIN + TONIC DIET METAL SIGN 0.022 0.471 8.9358
WOODEN PICTURE FRAME WHITE FINISH → WOODEN FRAME ANTIQUE WHITE 0.0378 0.5989 8.8756
WOODEN FRAME ANTIQUE WHITE → WOODEN PICTURE FRAME WHITE FINISH 0.0378 0.5601 8.8756

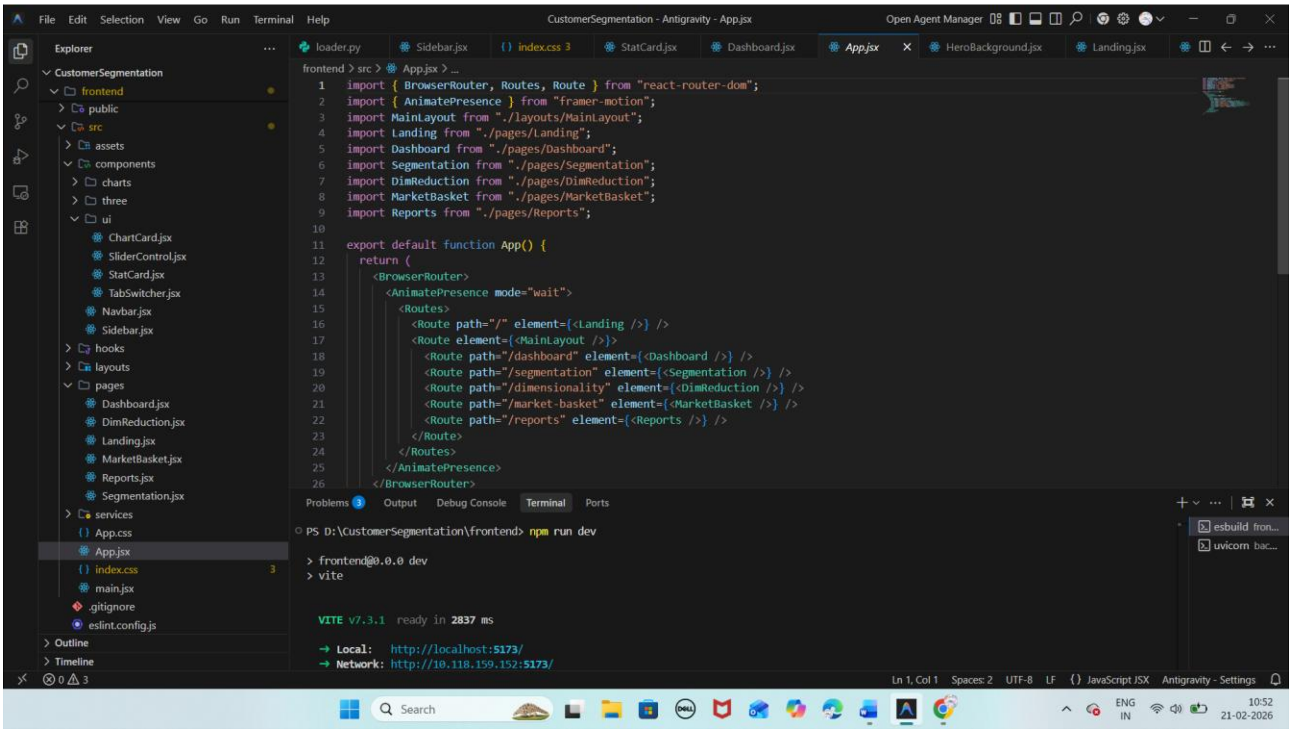

SegmentIQ — Customer Intelli
localhost:5173/market-basket
Gmail
YouTube
Maps
News
Translate
Adobe Acrobat
All Bookmarks
SegmentIQ
Intelligence Platform
Market Basket Analysis
UCI Online Retail II · Real ML Analysis
Backend: localhost:8000
Dashboard
Segmentation
Dim. Reduction
Market Basket
Reports
Product Co-occurrence Heatmap
How often pairs of products appear in the same basket (top 12×12)
60 TEATIME FAIRY CAKE
ASSORTED COLOUR BIRD O
BAKING SET 9 PIECE RET
HEART OF WICKER LARGE
HEART OF WICKER SMALL
HOME BUILDING BLOCK WO
JUMBO BAG RED RETROSPO
JUMBO SHOPPER VINTAGE
JUMBO STORAGE BAG SUKI
LUNCH BAG BLACK SKULL
LUNCH BAG CARS BLUE
LUNCH BAG SPACEBOY DES
Darker purple = higher co-occurrence in same basket
Search
ENG IN
10:50
21-02-2026

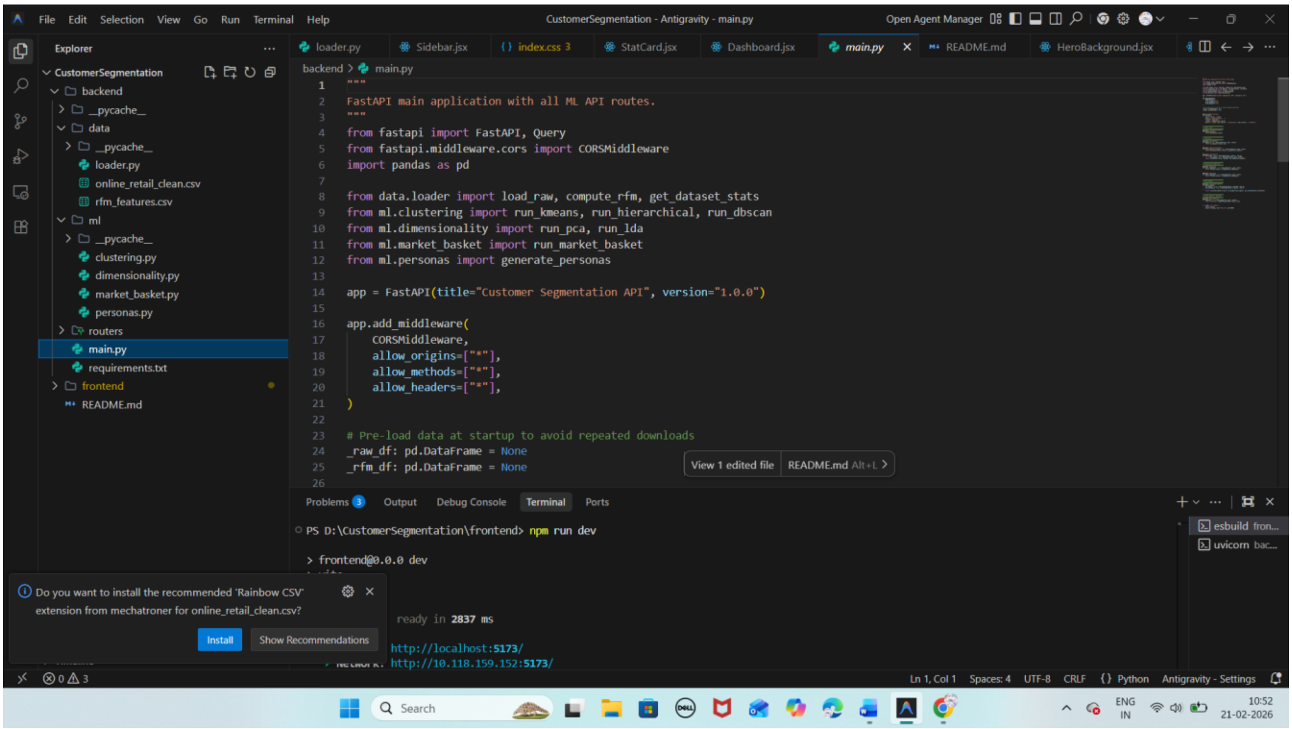

SegmentIQ — Customer Intelli
localhost:5173/market-basket
Gmail
YouTube
Maps
News
Translate
Adobe Acrobat
All Bookmarks
SegmentIQ
Intelligence Platform
Market Basket Analysis
UCI Online Retail II · Real ML Analysis
Backend: localhost:8000
Dashboard
Segmentation
Dim. Reduction
Market Basket
Reports
60 TEATIME FAIRY CAKE
ASSORTED COLOUR BIRD O
BAKING SET 9 PIECE RET
HEART OF WICKER LARGE
HEART OF WICKER SMALL
HOME BUILDING BLOCK WO
JUMBO BAG RED RETROSPO
JUMBO SHOPPER VINTAGE
JUMBO STORAGE BAG SUKI
LUNCH BAG BLACK SKULL
LUNCH BAG CARS BLUE
LUNCH BAG SPACEBOY DES
Darker purple = higher co-occurrence in same basket
Support
Fraction of baskets containing the itemset. Higher support = more common pattern.
supp(A→B) = P(A ∪ B)
Confidence
How often the rule is correct. P(B|A) — probability of B given A was purchased.
conf(A→B) = P(A ∪ B) / P(A)
Lift
Lift > 1 means items are bought together more than by chance. Key metric for actionable rules.
lift = conf / P(B)
Search
ENG IN
10:50
21-02-2026


SegmentIQ — Customer Intellig
localhost:5173/reports
Gmail
YouTube
Maps
News
Translate
Adobe Acrobat
All Bookmarks
SegmentIQ
Intelligence Platform
Cluster Reports & Insights
UCI Online Retail II · Real ML Analysis
Backend: localhost:8000
Dashboard
Segmentation
Dim. Reduction
Market Basket
Reports
Cluster Reports & Insights
Marketing personas derived from real K-Means cluster centroids
Number of Clusters (k)
4
2
8
High-Value Occasional
Cluster 0
20.5%
1,204 customers
Spend a lot but purchase infrequently.
Recency
Frequency
Monetary
Recency Score
Frequency
Monetary
Diversity
Basket Size
Campaign Strategy
Target with premium products and curated collections.
Loyal Customer
Cluster 1
21.4%
1,257 customers
Frequent buyers with consistent spend.
Campaign Strategy
Upsell higher-value products, ask for reviews and referrals.
Lost Customer
Cluster 2
32.1%
1,884 customers
Long inactive — haven't purchased in months.
Campaign Strategy
Reactivation campaigns with aggressive discounts.
Lost Customer
Cluster 3
26.1%
1,533 customers
Long inactive — haven't purchased in months.
Campaign Strategy
Reactivation campaigns with aggressive discounts.
Search
ENG
IN
10:50
21-02-2026


SegmentIQ — Customer Intellig
localhost:5173/reports
Gmail
YouTube
Maps
News
Translate
Adobe Acrobat
All Bookmarks
SegmentIQ
Intelligence Platform
Cluster Reports & Insights
UCI Online Retail II · Real ML Analysis
Backend: localhost:8000
Dashboard
Segmentation
Dim. Reduction
Market Basket
Reports
Business Intelligence Summary
Consolidated cluster report with revenue potential per segment
Persona | Customers | Share | Avg Recency | Avg Frequency | Avg Spend £ | Revenue Potential
High-Value Occasional | 1,204 | 20.5% | 27.5d | 3.5 | £1151 | £1385K
Loyal Customer | 1,257 | 21.4% | 33d | 18.7 | £10448 | £13133K
Lost Customer | 1,884 | 32.1% | 375.5d | 1.4 | £324 | £610K
Lost Customer | 1,533 | 26.1% | 262d | 4.3 | £1706 | £2616K
Total | 5,878 | 100% | | | | £17743K
Retain Champions
Invest in VIP loyalty programs and early-access campaigns for your highest-value customers. Churn here is costliest.
Win Back At-Risk
Trigger personalized re-engagement emails with time-limited discounts for customers who haven't purchased in 60+ days.
Upsell Loyals
Loyal customers are prime candidates for bundle offers and premium product recommendations to increase basket size.
Onboard New Customers
A structured onboarding journey with welcome offers and product discovery emails can convert new customers to loyals.
Search
ENG
IN
10:50
21-02-2026

FILE STRUCTURE:





Conclusion:

SegmentIQ demonstrates how unsupervised machine learning can transform raw e-commerce transaction data into structured, actionable business intelligence. By combining clustering algorithms, dimensionality reduction, and association rule mining within a full-stack interactive dashboard, the platform enables automated customer segmentation and strategic marketing decision-making.

The modular and scalable architecture allows integration with new datasets, additional machine learning models, and cloud deployment, making SegmentIQ a practical and extensible analytics solution for modern e-commerce businesses.