Pattern Recognition for Customer Behavior Analysis: A Study Using the UCI Online Retail II Dataset

Reporter: YUNG-CHUN LAN

Motivation

- To understand how customers behave based on their past purchases.
- To see how well classical PR methods can help us find different customer.

Overview

1. Data cleaning, aggregation, visualization
2. Data standardization, dimensionality reduction (PCA)
3. Unsupervised learning method (K-means)
4. Supervised learning method (compare LDA, KNN and Naïve Gaussian)

Raw Data: 541,910 rows, 8 columns
After Data Cleaning: 397,925 rows, 8 columns

Invoice	Invoice number (those starting with “C” indicate returns)
StockCode	Product code
Description	Product description
Quantity	Quantity purchased
InvoiceDate	Transaction date and time
Price	Unit price
CustomerID	Customer identifier
Country	Customer’s country

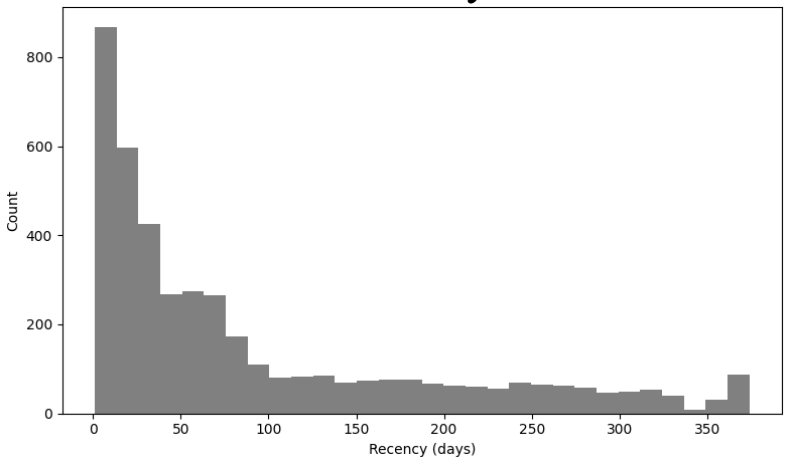
Invoice	Stockcode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom

Aggregation on Customer ID (4339 rows, 8 columns)

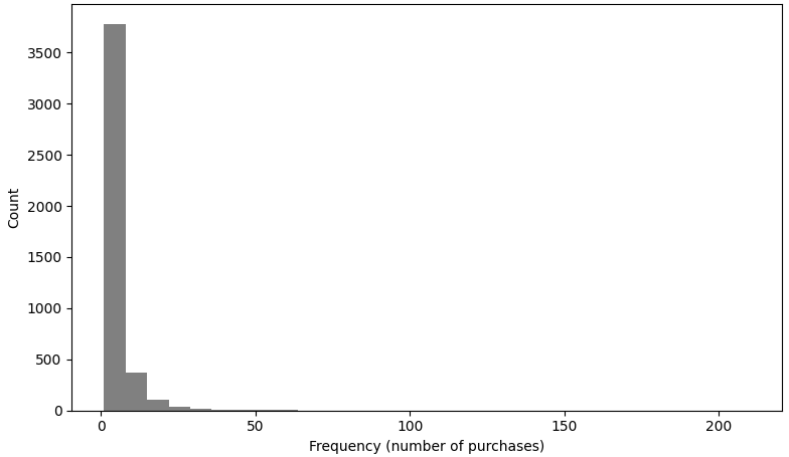
Feature	Description	Effect
Recency	Number of days since the last purchase	Customer activity level
Frequency	Number of purchases	High-frequency vs. low-frequency customer
Monetary	Total spending	Measuring customer value
AvgUnitPrice	Average unit price	Assessing spending level
UniqueItems	Number of product types purchased	Identifying diverse vs. specialized customer groups
AvgQuantPerOrder	Average quantity per order	Assessing bulk vs. non-bulk purchasing behavior

Visualization

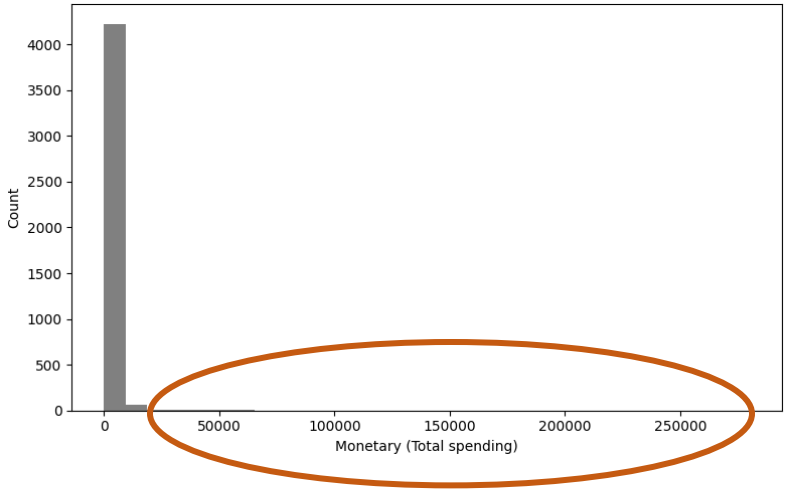
Recency



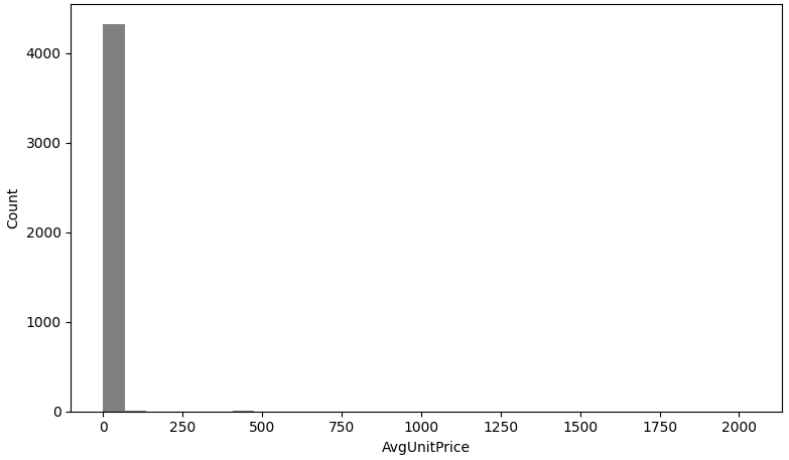
Frequency



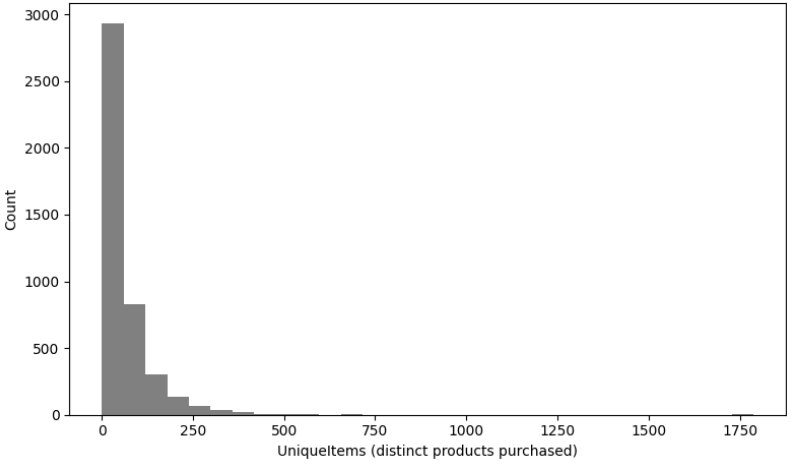
Monetary



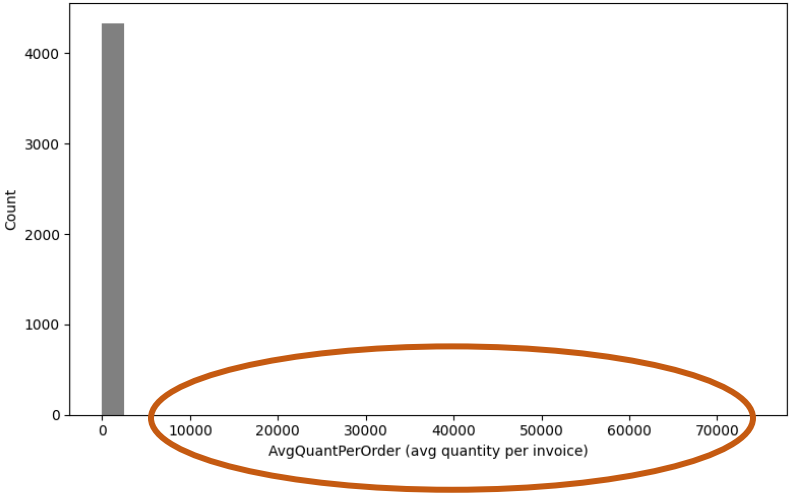
AvgUnitPrice

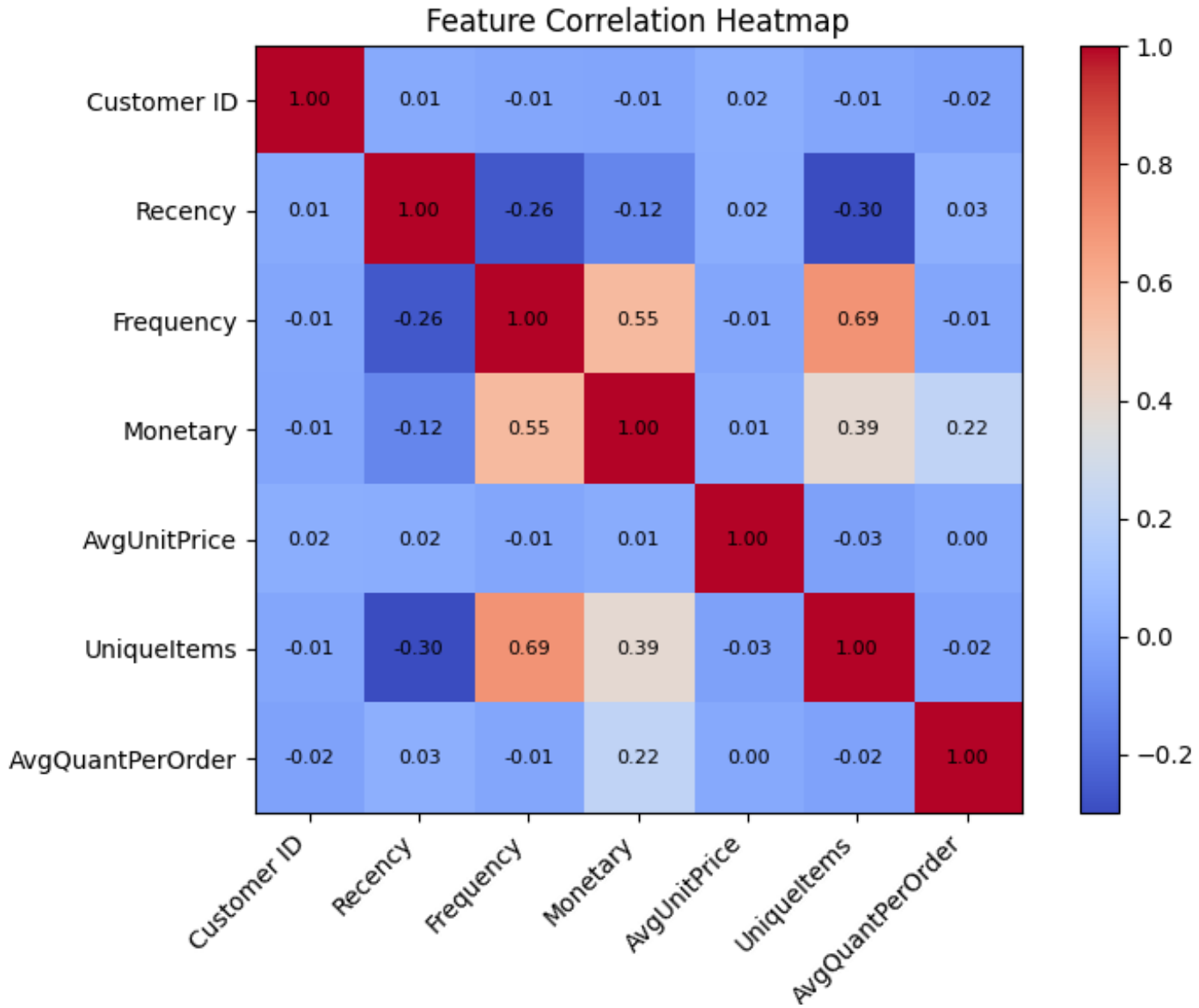


UniqueItems



AvgQuantPerOrder




Feature Correlation Heatmap
Customer ID
Recency
Frequency
Monetary
AvgUnitPrice
UniqueItems
AvgQuantPerOrder
1.00 0.01 -0.01 -0.01 0.02 -0.01 -0.02
0.01 1.00 -0.26 -0.12 0.02 -0.30 0.03
-0.01 -0.26 1.00 0.55 -0.01 0.69 -0.01
-0.01 -0.12 0.55 1.00 0.01 0.39 0.22
0.02 0.02 -0.01 0.01 1.00 -0.03 0.00
-0.01 -0.30 0.69 0.39 -0.03 1.00 -0.02
-0.02 0.03 -0.01 0.22 0.00 -0.02 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Standardization

$$z_i = \frac{x_i - \mu}{\sigma}$$

ID	Recency _scaled	Frequency _scaled	Monetary _scaled	AvgUnitPrice _scaled	UniqueItems _scaled	AvgQuantPerOrder _scaled
12346	2.33	-0.42	8.36	-0.10	-0.71	60.89
12347	-0.91	0.35	0.25	-0.05	0.49	-0.03
12348	-0.18	-0.04	-0.03	0.04	-0.46	0.02

PCA

	Recency_scaled	Frequency_scaled	Monetary_scaled	AvgUnitPrice_scaled	UniqueItems_scaled	AvgQuantPerOrder_scaled	Explained_Var ratio
PC1	**-0.308**	**0.595**	**0.484**	-0.017	**0.559**	0.066	0.37
PC2	0.362	-0.052	0.379	0.154	-0.165	**0.820**	0.19
PC3	-0.009	0.033	-0.005	**0.984**	0.014	-0.173	0.17
PC4	0.851	0.222	0.192	-0.073	0.113	-0.414	0.14
PC5	-0.223	-0.110	0.701	-0.047	-0.574	-0.339	0.09
PC6	-0.015	0.762	-0.307	-0.005	-0.563	0.084	0.04

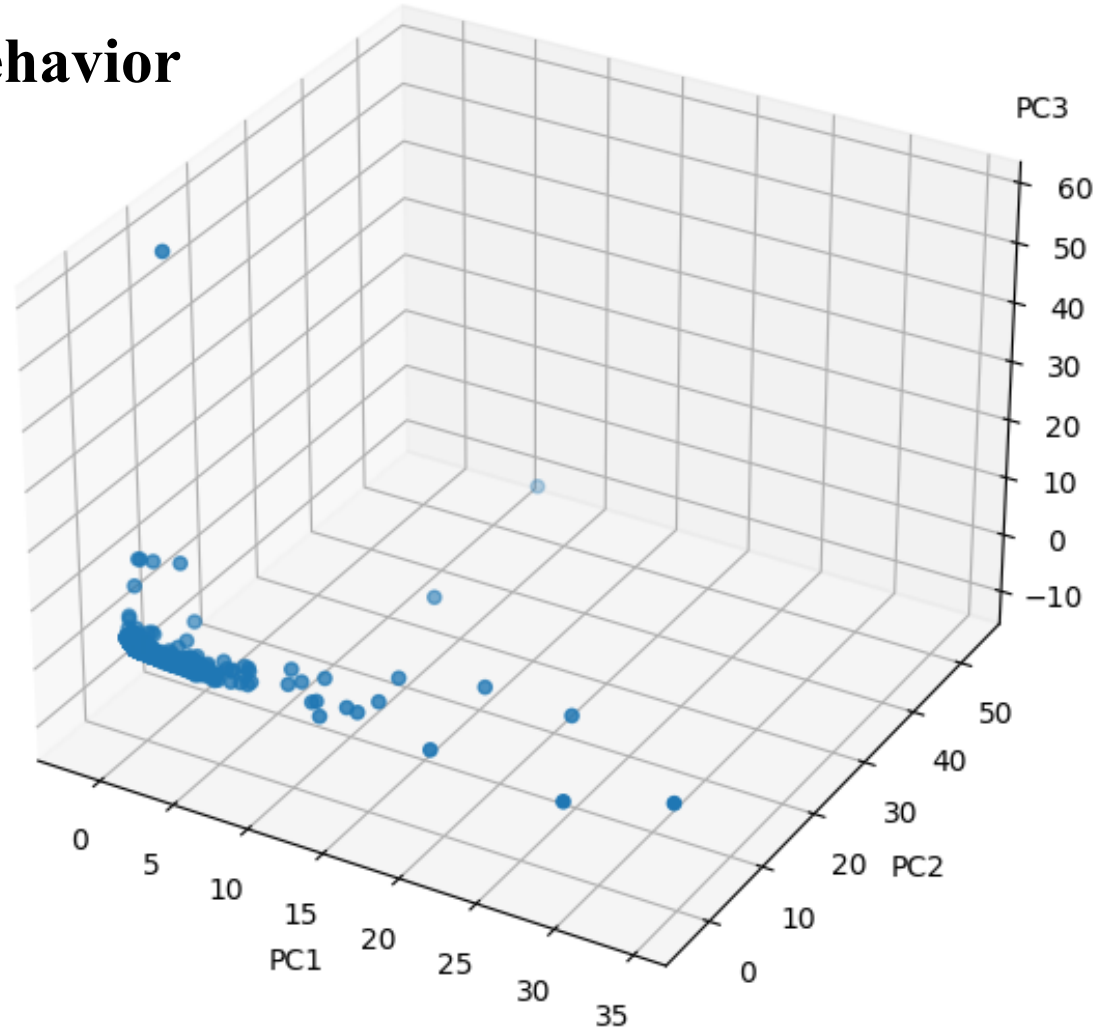
73%

- PC1 (37.3%): Customer Value and Activity
- PC2 (18.6%): Bulk or Wholesale Purchasing Behavior
- PC3 (16.6%): Unit Price Preference

PC1 : Customer Value and Activity

PC2 : Bulk or Wholesale Purchasing Behavior

PC3 : Unit Price Preference



K-means Clustering

Find:

$$\min J = \sum_{i=1}^{n} ||x_i - \mu_{c_i}||^2$$

Repeat two simple steps:

- Step 1 (Assign points)

$$c_i = \underset{j \in \{1, \dots, k\}}{\operatorname{argmin}} ||x_i - u_j||^2$$

- Step 2 (Update centers)

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i \ , C_k = \{x_i : c_i = k\}$$

Until:

$$\mu_k^{(t+1)} = \mu_k^{(t)} \ \forall \ k \ \ or \ \ J^{(t+1)} - J^{(t)} < \varepsilon$$

Silhouette Score

$$S = \frac{1}{n}\sum_{i=1}^{n} s(i)$$

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

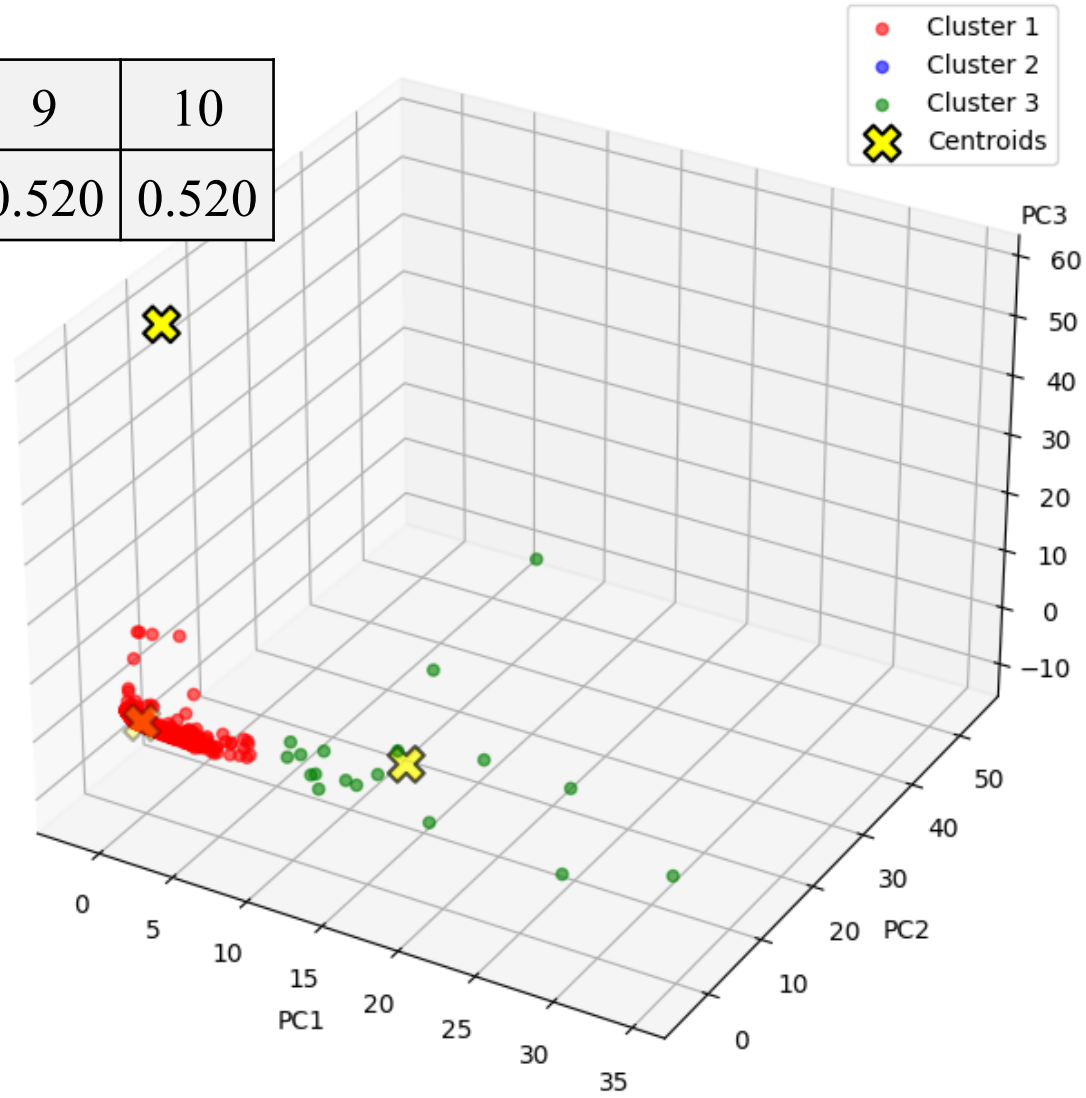
$a(i)$: the average distance from point i to other points in the same cluster.

$b(i)$: the average distance from point i to the closest different cluster.

K-means after PCA

k	2	3	4	5	6	7	8	9	10
score	0.932	0.924	0.925	0.563	0.510	0.514	0.518	0.520	0.520

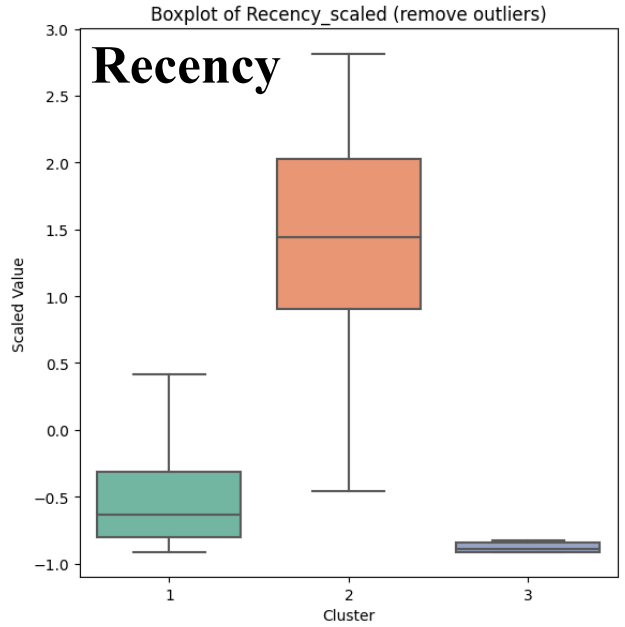
Cluster ID	1	2	3
Quantity	4320	1	18

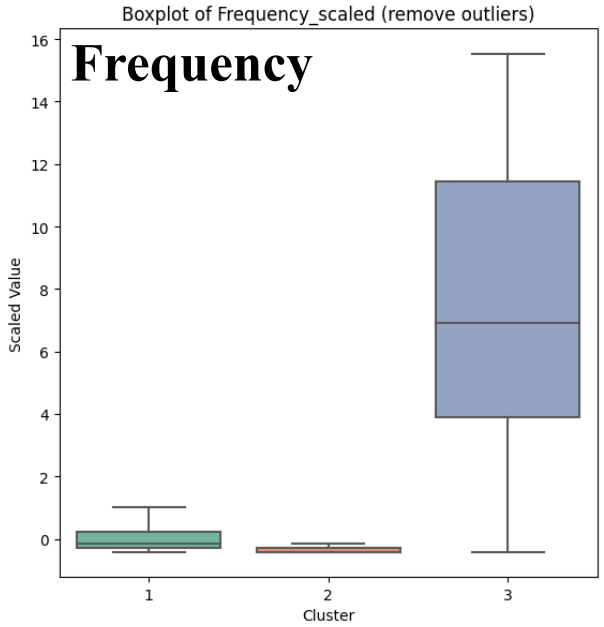


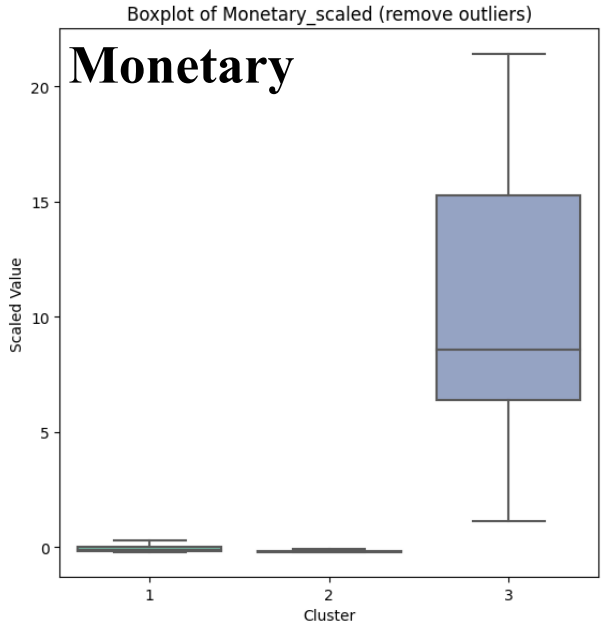
K-means without PCA

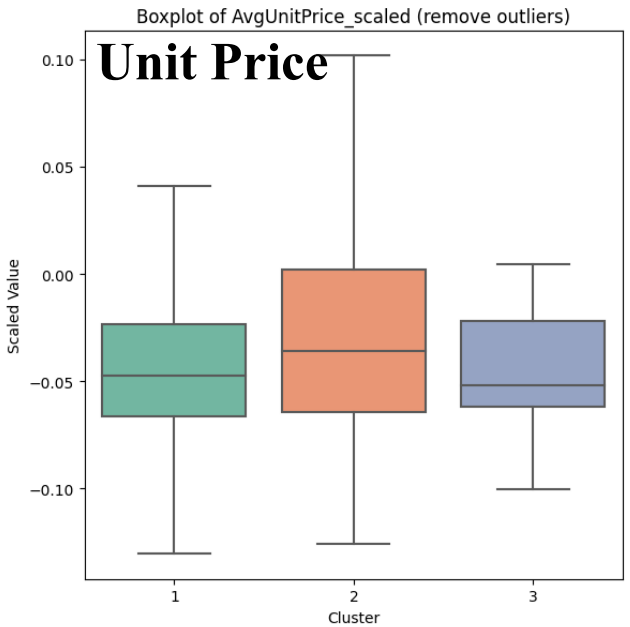
k	2	3	4	5	6	7	8	9	10
score	0.446	0.470	0.466	0.466	0.466	0.510	0.470	0.502	0.399

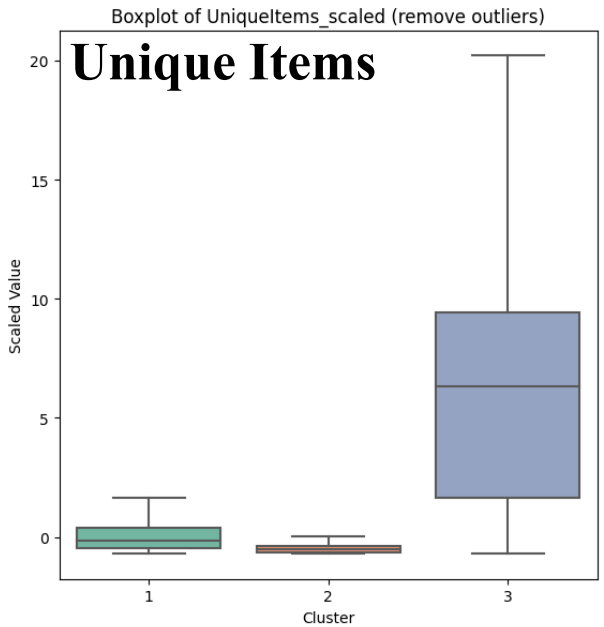
Cluster ID	1	2	3
Quantity	3180	1140	18

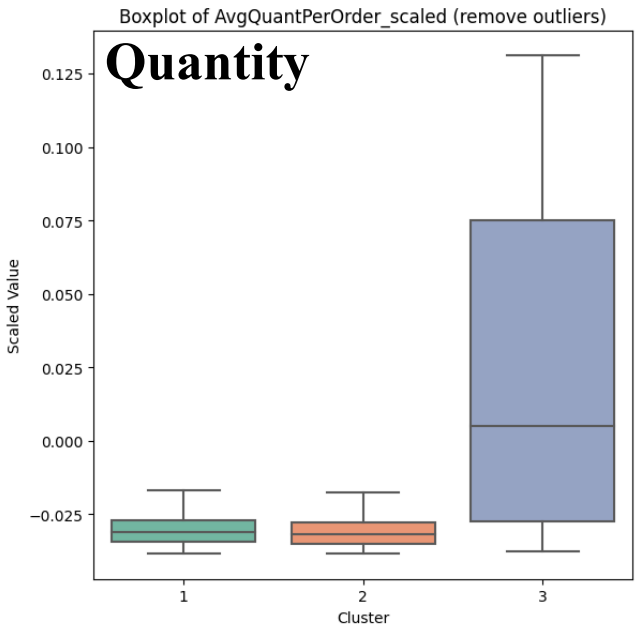

Boxplot of Recency_scaled (remove outliers)
Recency
Scaled Value
Cluster

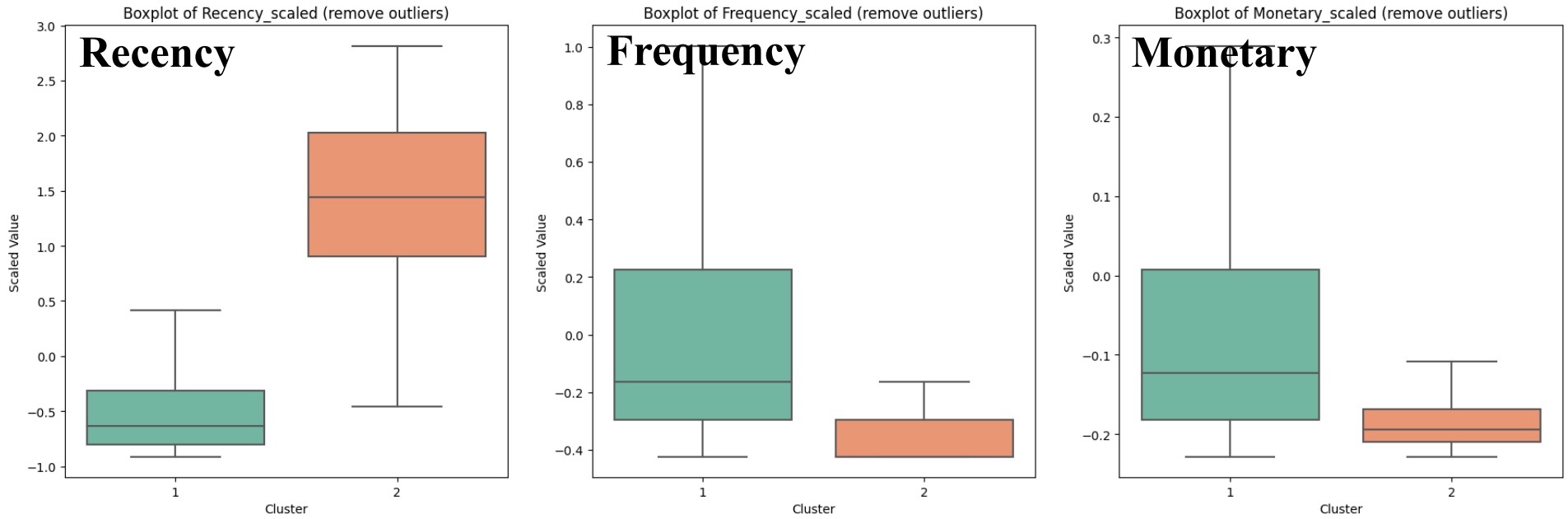

Boxplot of Frequency_scaled (remove outliers)
Frequency
Scaled Value
Cluster

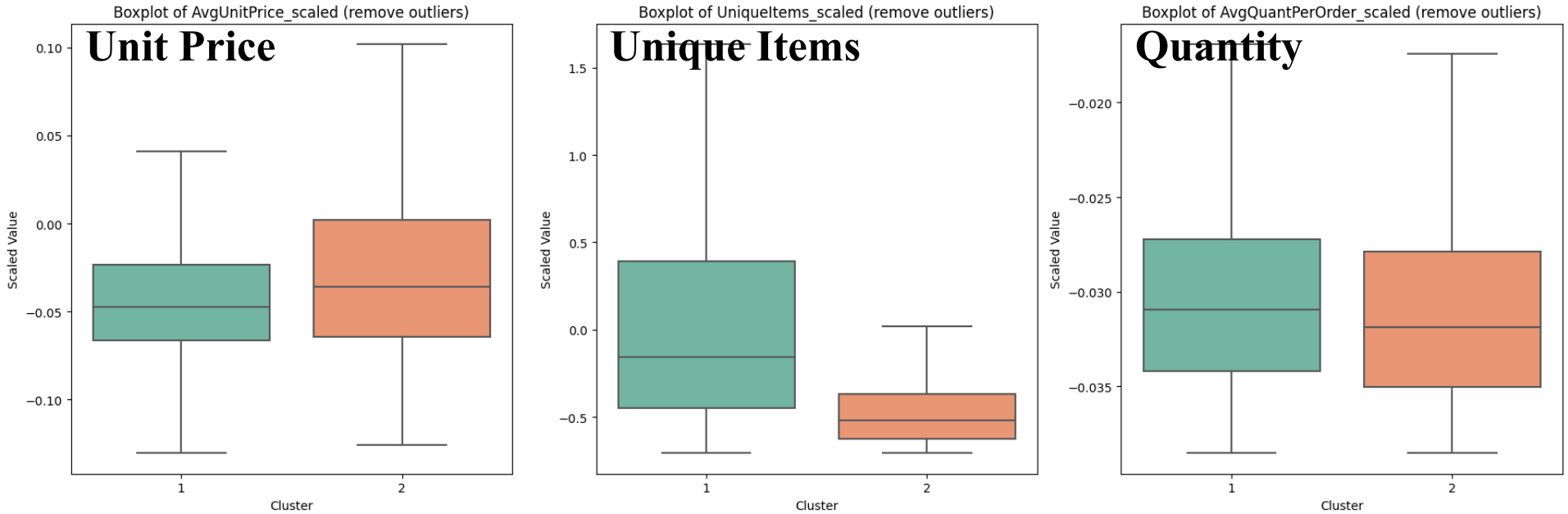

Boxplot of Monetary_scaled (remove outliers)
Monetary
Scaled Value
Cluster


Boxplot of AvgUnitPrice_scaled (remove outliers)
Unit Price
Scaled Value
Cluster


Boxplot of UniqueItems_scaled (remove outliers)
Unique Items
Scaled Value
Cluster


Boxplot of AvgQuantPerOrder_scaled (remove outliers)
Quantity
Scaled Value
Cluster


Boxplot of Recency_scaled (remove outliers)
Recency
Scaled Value
Cluster
Boxplot of Frequency_scaled (remove outliers)
Frequency
Scaled Value
Cluster
Boxplot of Monetary_scaled (remove outliers)
Monetary
Scaled Value
Cluster


Boxplot of AvgUnitPrice_scaled (remove outliers)
Unit Price
Scaled Value
Cluster
Boxplot of UniqueItems_scaled (remove outliers)
Unique Items
Scaled Value
Cluster
Boxplot of AvgQuantPerOrder_scaled (remove outliers)
Quantity
Scaled Value
Cluster

Label

- C1: Regular shoppers
- C2: Inactive low-value shoppers
- C3: High-value bulk buyers

Clustering Quality

- k=3
- Silhouette score = 0.470
- Sizes (3181, 1140, 18)
- The boxplots show clear differences across three groups

LDA

Assumption:

$$p(\vec{x}|w_i) \sim N(\vec{\mu}_i, \Sigma)$$

Discrimination function:

$$g_i(\vec{x}) = -\frac{1}{2}(\vec{x} - \vec{\mu}_i)^T \Sigma^{-1}(\vec{x} - \vec{\mu}_i) + \ln P(\omega_i)$$

$$= \vec{w}_i^T \vec{x} + w_{i0}$$

$$where\ \vec{w}_i = \Sigma^{-1}\vec{\mu}_i, \qquad w_{i0} = -\frac{1}{2}\vec{\mu}_i^T \Sigma^{-1}\vec{\mu}_i + \ln P(\omega_i)$$

Decision boundary:

$$\vec{w}^T(\vec{x} - \vec{x}_0) = \vec{0},$$

$$where\ \vec{w} = \Sigma^{-1}(\vec{\mu}_i - \vec{\mu}_j), \vec{x}_0 = \frac{1}{2}(\vec{\mu}_i + \vec{\mu}_j) - \frac{\ln[\frac{P(\omega_i)}{P(\omega_j)}]}{(\vec{\mu}_i - \vec{\mu}_j)^T \Sigma^{-1}(\vec{\mu}_i - \vec{\mu}_j)}(\vec{\mu}_i - \vec{\mu}_j)$$

KNN

$$D = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_n, y_n)\}$$

$$where\ \vec{x}_i \in \mathbb{R}^d : d - dimensional\ feature\ vector,$$

$$y_i \in \{\omega_1, \dots, \omega_C\} : class\ label$$

if $N_k(x)$ denotes the set of the k nearest neighbors of $\vec{x}$,

then:

$$\hat{\omega}(x) = \underset{\omega_c}{argmax} \sum_{x_j \in N_k(x)} I_{\{y_j = \omega_c\}}$$

Naïve Gaussian

Assumption: All features are conditionally independent within the same classes

$$p(x|\omega_i) = \prod_{j=1}^{d} p(x_j|\omega_i)$$

Each feature follows a one-dimensional normal distribution:

$$x_j|\omega_i \sim N(\mu_{ij}, \sigma_{ij}^2)$$

Discriminant function:

$$g_i(x) = \sum_{j=1}^{d} [-\frac{1}{2}\ln(2\pi\sigma_{ij}^2) - \frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2}]$$

Concepts & Evaluation Basis

- Correctly identifying the High-Value Bulk Buyers (C3) is the top priority, but C3 is minority.

➢ Confusion matrix ($Recall_i = \frac{aii}{ai1+ai2+ai3}$, $Precision_i = \frac{aii}{a1i+a2i+a3i}$)

Actual \ Pred	Pred C1	Pred C2	Pred C3
Actual C1	a11	a12	a13
Actual C2	a21	a22	a23
Actual C3	a31	a32	a33

➢ Macro-average F1 score

$$F1_{macro} = \frac{1}{C}\sum_{i=1}^{C} F1_i \ , where\ F1_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i}$$

➢ Balanced accuracy score

$$Balanced\ Accuracy = \frac{1}{C}\sum_{i=1}^{C} Recall_i$$

Results of LDA (60% training 40% testing)

- Confusion matrix:

$$\begin{bmatrix} 1269 & 4 & 0 \\ 42 & 414 & 0 \\ 0 & 0 & 7 \end{bmatrix}$$

- Classification report:

	precision	recall	F1-score	support
C1	0.97	1.00	0.98	1273
C2	0.99	0.91	0.95	456
C3	1.00	1.00	1.00	7
Acc.				0.97
Macro Avg. F1.				0.98
Balanced Acc.				0.97

Results of KNN (60% training 40% testing)

- Confusion matrix:

$$\begin{bmatrix} 1268 & 5 & 0 \\ 6 & 450 & 0 \\ 2 & 0 & 5 \end{bmatrix}$$

- Classification report:

	precision	recall	F1-score	support
C1	0.99	1.00	0.99	1273
C2	0.99	0.91	0.99	456
C3	1.00	0.71	0.83	7
Acc.				0.99
Macro Avg. F1.				0.94
Balanced Acc.				0.90

Results of Naïve Gaussian (60% training 40% testing)

- Confusion matrix:

$$\begin{bmatrix} 1204 & 45 & 24 \\ 10 & 446 & 0 \\ 1 & 0 & 6 \end{bmatrix}$$

- Classification report:

	precision	recall	F1-score	support
C1	0.99	0.95	0.97	1273
C2	0.91	0.98	0.94	456
C3	0.2	0.86	0.32	7
Acc.				0.95
Macro Avg. F1.				0.74
Balanced Acc.				0.93

Discussion: (LDA > KNN > Naïve Gaussian)

1. LDA

The three clusters exhibit strong directional differences that make them close to linearly separable.

2. KNN

Can easily be misclassified if a data close to another class in certain dimensions

3. Naïve Gaussian

- sensitive to dimensions where samples appear close to other classes.
- the assumption that all features are independent within a class is unrealistic

Conclusion:

This case shows that when the dataset is:

1. The number of features is limited
2. Customer behaviors exhibit clear directional differences

Simple and interpretable linear models (such as LDA) can accurately identify high-value customers, without necessarily relying on complex models.