Wholesale Customer Segmentation: Identifying Purchasing Patterns with Clustering

Miguel Franco

https://github.com/MFranco2000/effective-octo-palm-tree

Problem Statement: Understanding Customer Segmentation

Businesses struggle to group customers effectively based on purchasing behavior.

Traditional methods rely on manual categorization, which is inefficient.

The goal: Use clustering techniques to identify natural customer segments based on spending patterns.

These insights can help optimize marketing, inventory management, and pricing strategies.

The Data: Wholesale Customers Dataset

Source: UCI Machine Learning Repository.

440 customers, each representing a wholesale client.

Six spending categories: Fresh, Milk, Grocery, Frozen, Detergents_Paper, and Delicatessen.

No predefined labels—ideal for unsupervised learning.

Cardoso, M. (2013). Wholesale customers [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5030X.

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	2	3	12669	9656	7561	214	2674	1338
1	2	3	7057	9810	9568	1762	3293	1776
2	2	3	6353	8808	7684	2405	3516	7844
3	1	3	13265	1196	4221	6404	507	1788
4	2	3	22615	5410	7198	3915	1777	5185

Data Quality and Initial Insights

Summary statistics reveal different spending patterns.

No missing values detected—data is ready for analysis.

Some spending categories have high variance and potential outliers.

Summary Statistics:

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	1.322727	2.543182	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	0.468052	0.774272	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	1.000000	1.000000	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	1.000000	2.000000	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	1.000000	3.000000	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	2.000000	3.000000	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	2.000000	3.000000	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

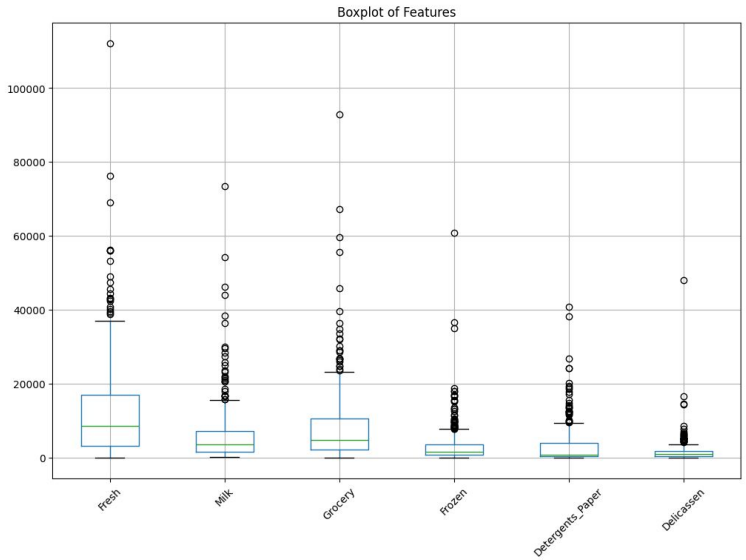
```
Missing Values:
Channel             0
Region              0
Fresh               0
Milk                0
Grocery             0
Frozen              0
Detergents_Paper    0
Delicassen          0
dtype: int64
```

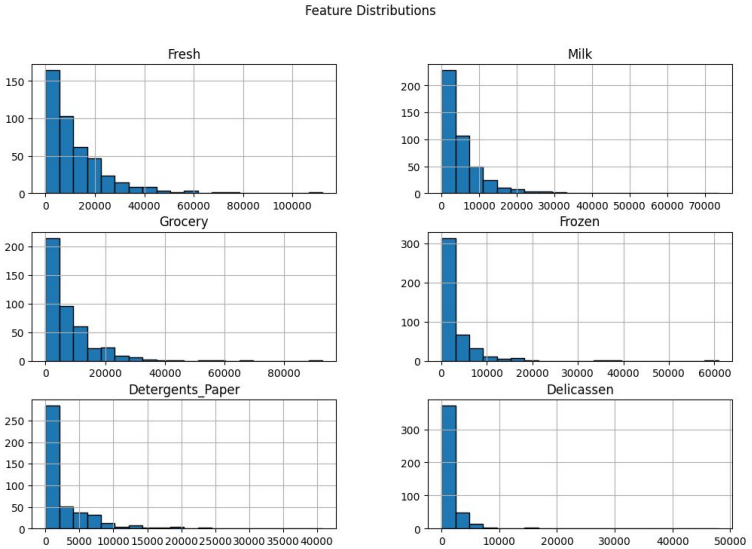
Understanding Feature Distributions

Histograms reveal skewed distributions in certain spending categories.

Some customers spend significantly more than others.

Standardization may be required for better clustering results.



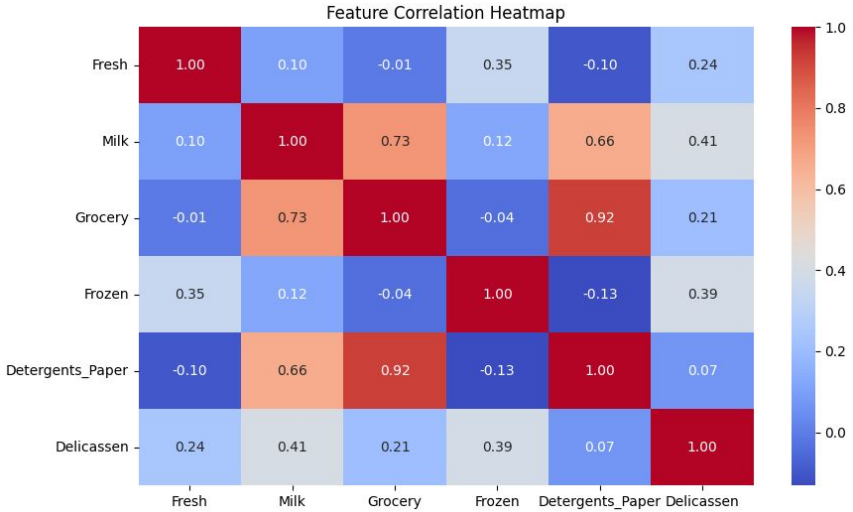


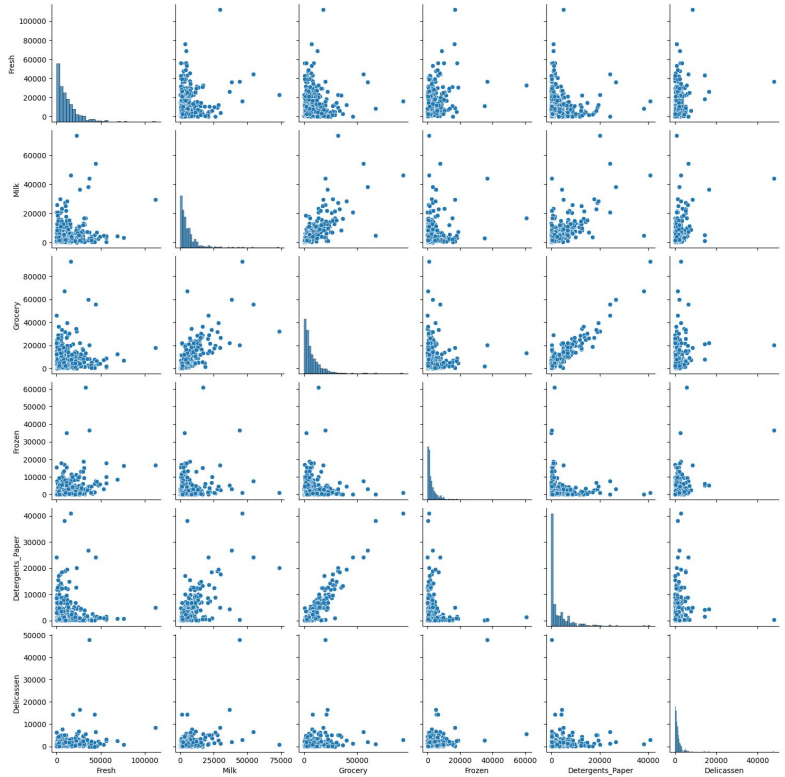
Feature Relationships & Correlation

A pair plot shows relationships between spending categories.

Correlation heatmap reveals strong correlations: Grocery & Detergents_Paper.

Understanding these relationships helps us decide which clustering methods may work best.



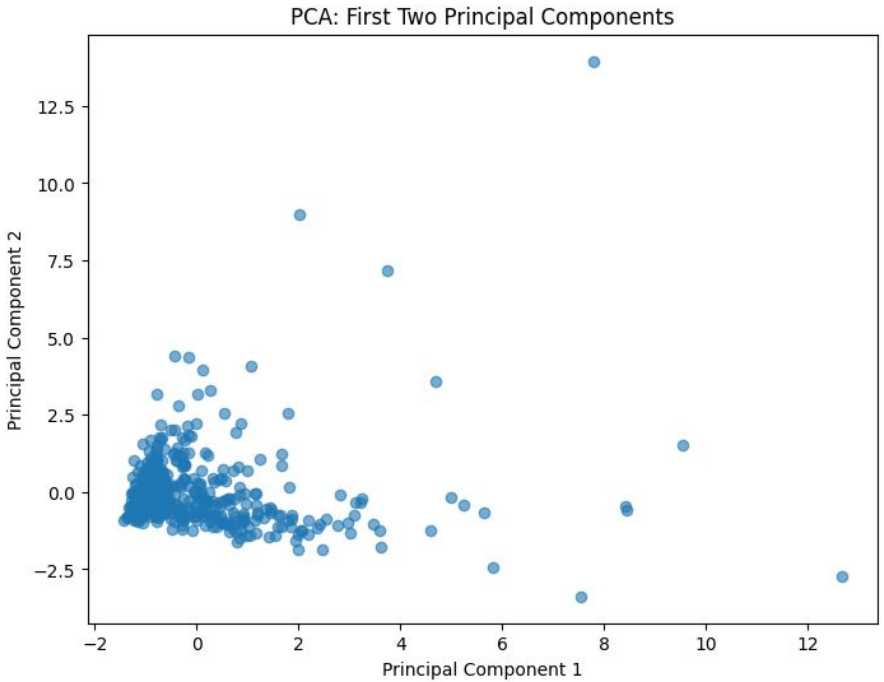


Dimensionality Reduction for Better Visualization

PCA helps reduce complexity while retaining key patterns.

The first two principal components explain most of the variance.

Helps in visualizing clusters before applying models.



Clustering Approaches

We apply three clustering methods to find customer segments:

- K-Means: Assigns customers to the nearest cluster center.
- Hierarchical Clustering: Groups customers in a tree-like structure.
- DBSCAN: Identifies dense regions and noise points.

Different methods handle data differently, leading to varied results.

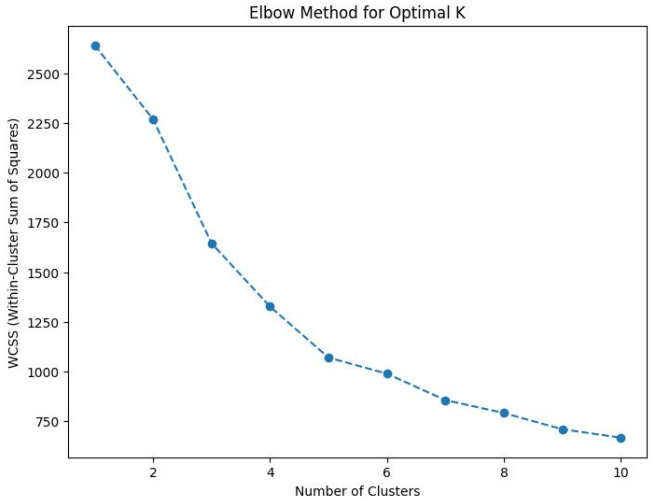
K-Means: Assigning Customers to Clusters

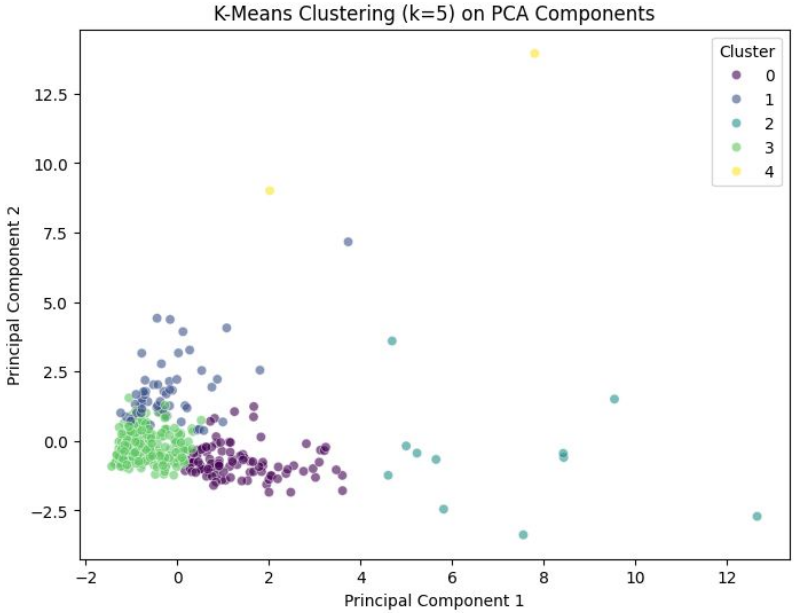
Requires choosing k (number of clusters) beforehand.

Used the Elbow Method to find the optimal k.

Silhouette score helps assess clustering quality.

Clusters formed around spending patterns.





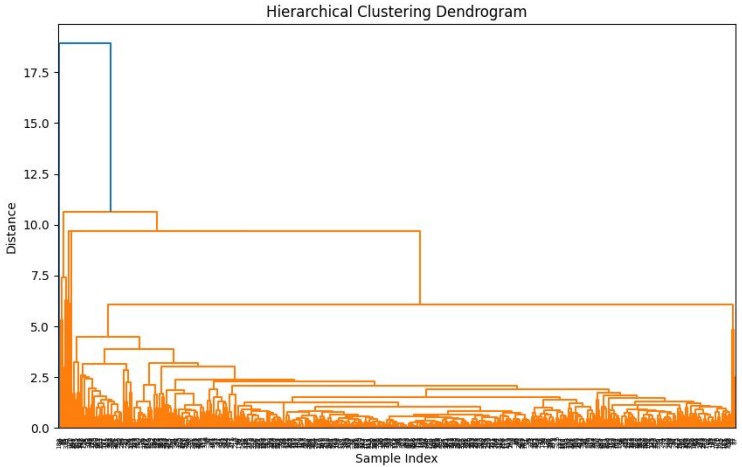
Hierarchical Clustering: Building a Customer Hierarchy

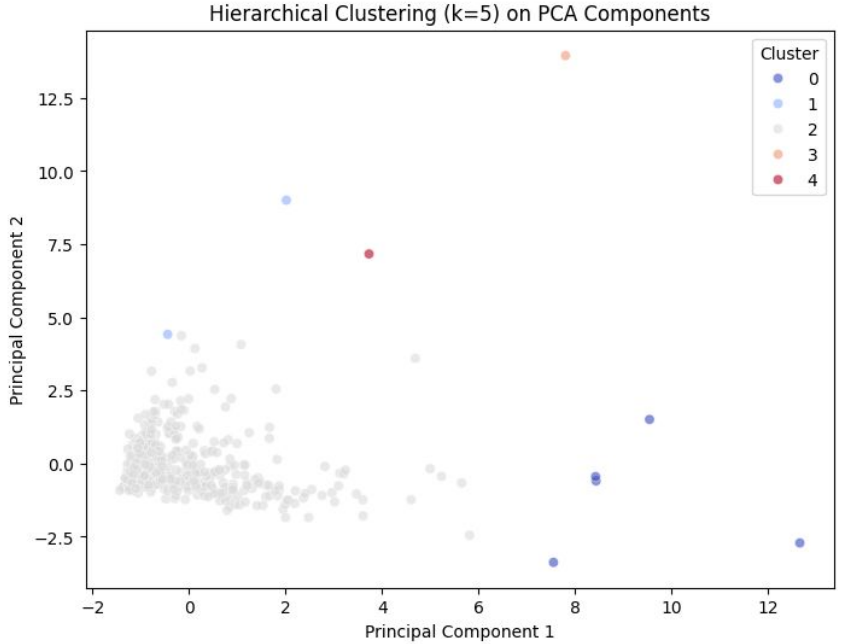
Does not require specifying k beforehand.

Uses linkage methods to build cluster hierarchies.

Dendrogram helps visualize relationships between customers.

Achieved strong silhouette scores.





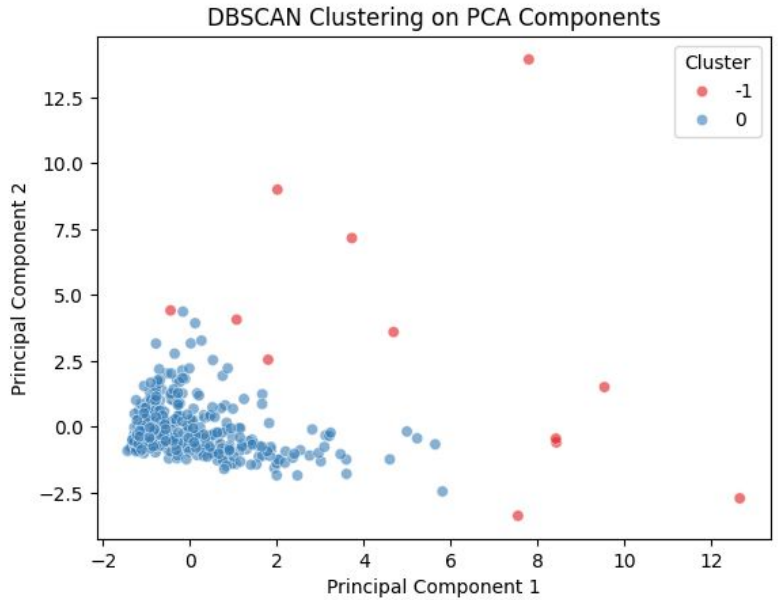
DBSCAN: Discovering Natural Customer Groups

Unlike K-Means, doesn't assume clusters are spherical.

Can identify noise points that don't fit into any cluster.

Required hyperparameter tuning (eps & min_samples) for best performance.

Achieved the highest silhouette score among all methods.



Which Clustering Method Works Best?

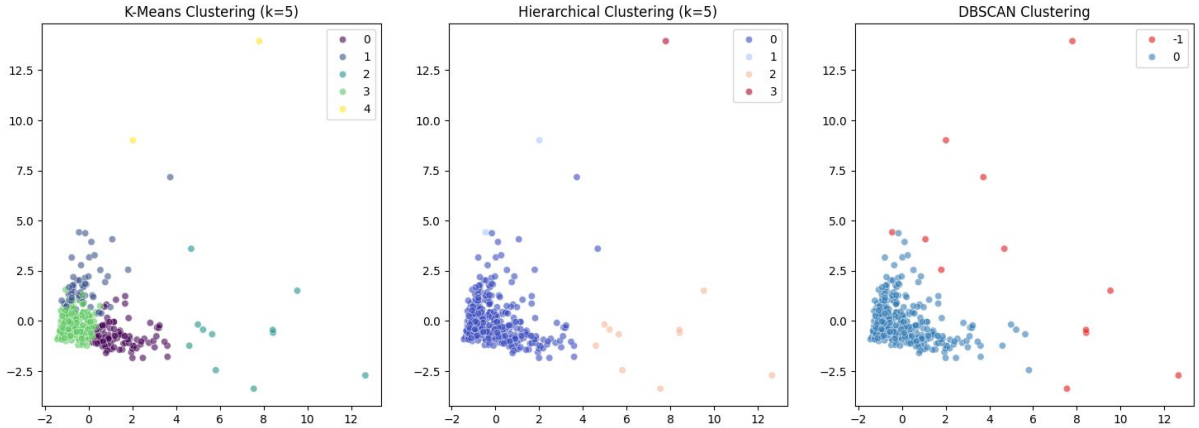
Silhouette Scores:

- K-Means (k=5): 0.37
- Hierarchical (k=5, average linkage): 0.74
- DBSCAN (best eps): 0.75

DBSCAN performed best, capturing natural groups and noise points.

Hierarchical clustering also provided strong results, while K-Means was more rigid.

Different methods serve different purposes based on business needs.



Key Takeaways & Business Applications

Customer segmentation is crucial for optimizing marketing and supply chain strategies.

DBSCAN provided the best clustering but requires careful tuning.

Hierarchical clustering is useful for understanding customer relationships.
K-Means is simple but less flexible.

Future work: Exploring hybrid clustering approaches or incorporating additional features like purchase frequency.