REPORT

Clustering Analysis

Dataset: Wholesale Customer Dataset

Name: Pallak Sinha

Summary

An Overview of Approach & Key Findings

The analysis examines the Wholesale Customers dataset with the goal of identifying natural customer segments.

The workflow begins with thorough exploratory data analysis, including summary statistics, boxplots, correlation study, and outlier assessment. Due to high skewness in spending variables, log transformation was applied, followed by standardization to ensure all features contributed equally to distance calculations.

Feature engineering was also performed to derive meaningful variables such as *Total Spend, Retail%, and Perishable%*, providing deeper insight into purchasing patterns.

Principal Component Analysis (PCA) was then used to reduce dimensionality and identify the major drivers of variance, revealing two dominant spending dimensions: retail-oriented (Grocery, Milk, Detergents) and perishable-oriented (Fresh, Frozen, Delicassen).

Using these transformed and engineered features, KMeans clustering was implemented and evaluated using the Elbow Method and Silhouette Score, both of which indicated three optimal clusters. PCA scatterplots further verified the separability of these clusters.

The resulting segments showed clear real-world business meaning:

-Cluster 0 consisted of small-to-medium retail buyers with high retail purchases;

-Cluster 1 represented HoReCa (Hotel–Restaurant–Café) customers with high perishable demand; and

-Cluster 2 captured a small but extremely high-value and high-volume group of large wholesalers or retail chains.

The cluster profiles reflect realistic customer behaviours, and the analysis confirms that the chosen segmentation is both stable and meaningful for business decision-making.

Part 1

Data Exploration and Preparation with Visualizations

1.1 Dataset Overview

The Wholesale Customers dataset contains annual spending of 440 customers across six product categories: Fresh, Milk, Grocery, Frozen, Detergents_Paper, and Delicassen. In addition, two categorical attributes—Channel (HoReCa vs Retail) and Region—provide contextual information about the type and location of each buyer.

Using .head() and .info() data's variables and other pertinent information was examined.

Data auditing was done to find any missing values, duplicate entries and unique value counts in each variable.

1.2 Exploratory Data Analysis (EDA)

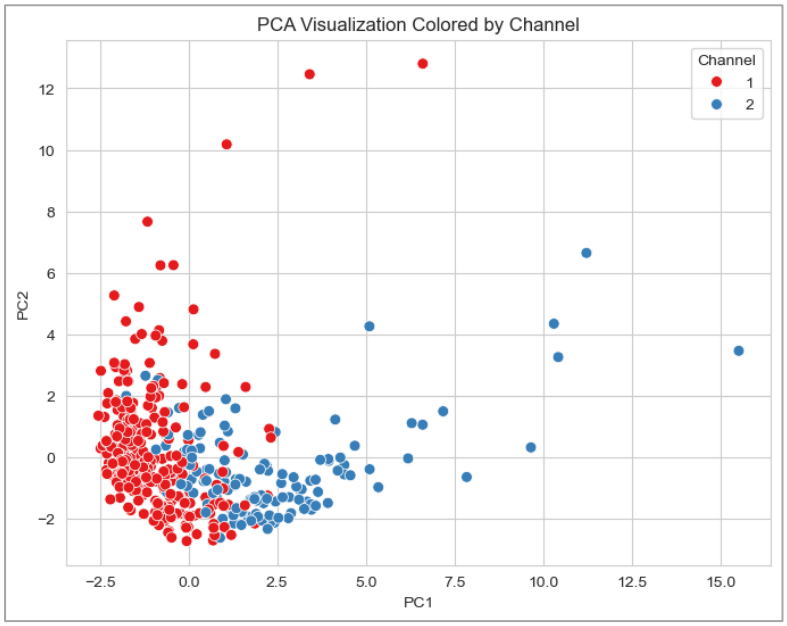
EDA was performed to understand distributions, relationships, and potential anomalies in the data.

Initial inspection using summary statistics revealed positive skewness in all expenditure variables and large variability across categories, indicating the presence of outliers and unequal scaling.

Key visualizations included:

- Boxplots for each spending category → revealed heavy right-skew and several extreme high-spending customers.
- Histogram → confirmed non-normal distributions typical for financial spending datasets.
- Correlation heatmap → showed strong correlation between Grocery, Milk, and Detergents_Paper, suggesting a retail-oriented spending pattern, while Fresh, Frozen, and Delicassen aligned with perishable-based customers (HoReCa) had negative correlation with Detergents_Paper implying that retail based customers would also buy fresh-frozen items but HoReCa customers only bought perishable products.
- PCA scatter (initial) → indicated natural separation between low-volume and high-volume buyers as well as highlighted the outlying unpredictable buyers.

These observations justified the need for transformation and scaling before clustering.



PCA Scatterplot showing data segmented on the basis of Channel

1.3 Data Cleaning and Outlier Handling

Given the skewed distributions and extreme values, log transformation was applied to all expenditure variables. This significantly reduced skewness, pulled extreme values closer to the core distribution, and improved symmetry for PCA and distance-based clustering. No rows were removed, instead, the transformation ensured their influence remained controlled.

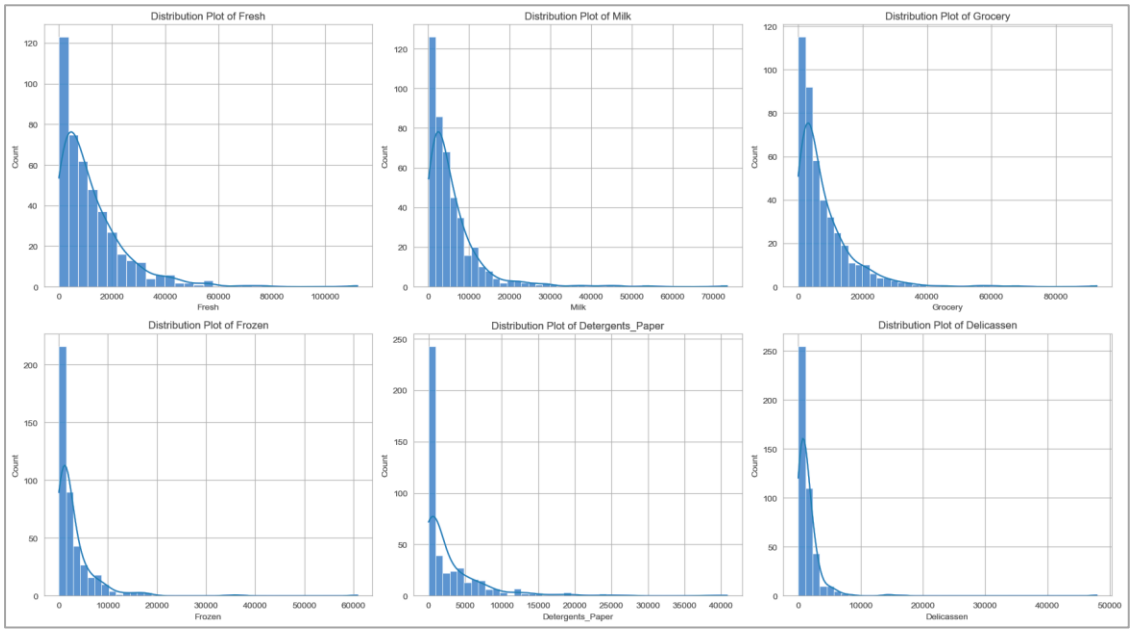
1.4 Feature Engineering

To capture more meaningful behavioural patterns, new features were constructed:

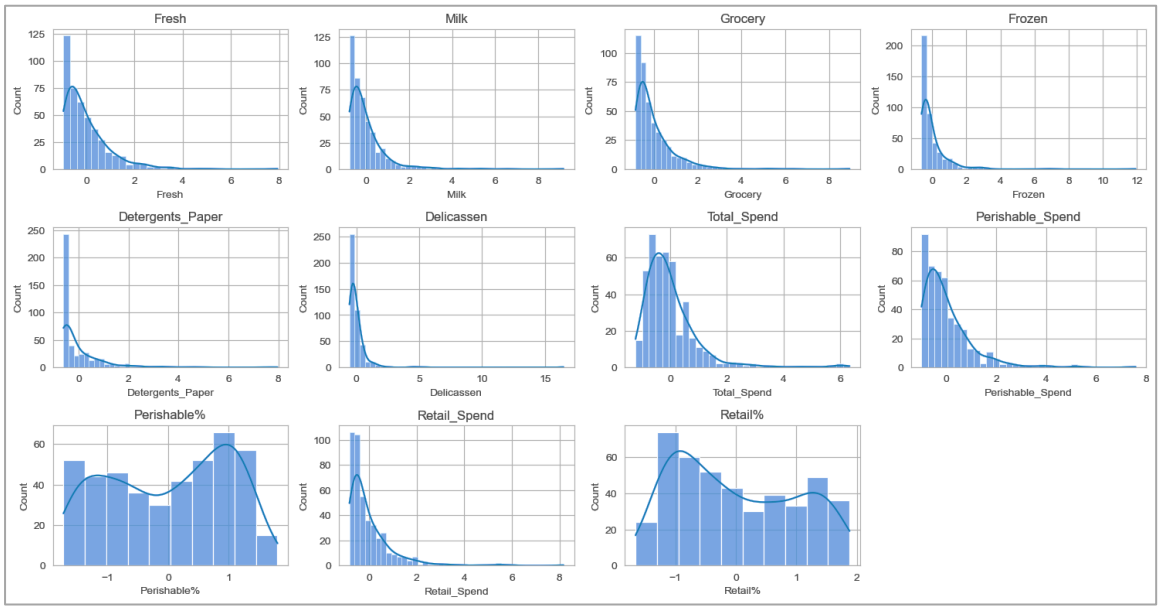
- Total_Spend = overall purchasing intensity using sum
- Retail% = (Grocery + Milk + Detergents_Paper) / Total_Spend
- Perishable% = (Fresh + Frozen + Delicassen) / Total_Spend

1.5 Scaling and Data Preparation for Modelling

Since clustering algorithms like KMeans are sensitive to feature magnitude, all numerical features—including engineered variables—were standardized using StandardScaler(). This ensured that no single variable dominated the distance calculations. The final pre-processed dataset was then passed into PCA and clustering models.



Initial Distribution of Numerical Features (Spending variables)



After transformation and scaling, all the distributions have mean=0 and unit variance

Part 2

Model Development, Comparison, Validation, and Diagnostics

2.1 Model Objective and Approach

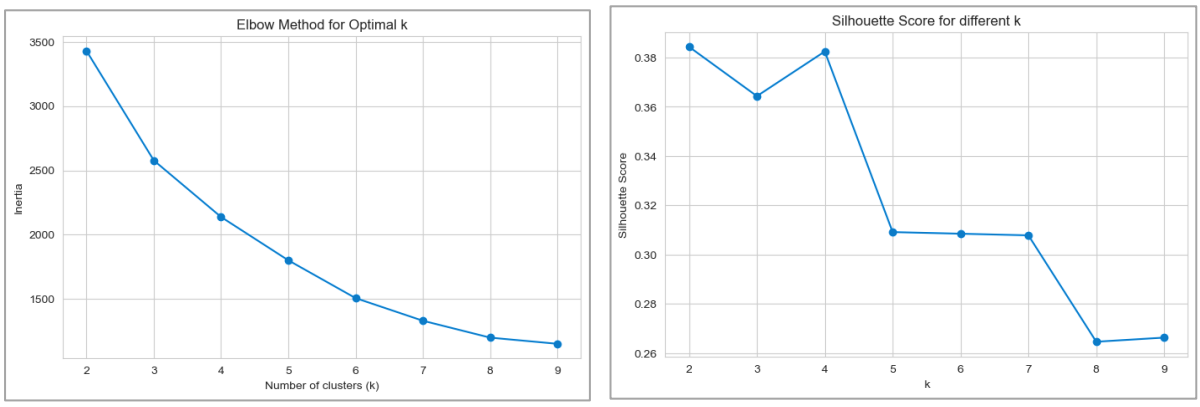
The objective of the modeling phase was to identify natural customer groupings within the Wholesale Customers dataset using unsupervised learning. Because all variables are continuous and represent annual expenditures, KMeans clustering was selected as the primary model. Before model training, all features—including engineered variables (Total_Spend, Retail%, Perishable%)—were standardized to remove scale effects, ensuring equal contribution to distance calculations.

2.2 Selecting the Optimal Number of Clusters

Two internal validation techniques were used to determine the appropriate number of clusters:

- Elbow Method: The inertia curve displayed a clear bend at k = 3, indicating diminishing reductions in within-cluster variance beyond this point.
- Silhouette Analysis: Silhouette scores peaked at k = 3, confirming that this configuration provided the best balance between cluster compactness and separation.

Both methods independently supported the selection of 3 clusters, making it the optimal solution for the dataset.



Figures showing elbow method and silhouette method output

2.3 KMeans Model Training

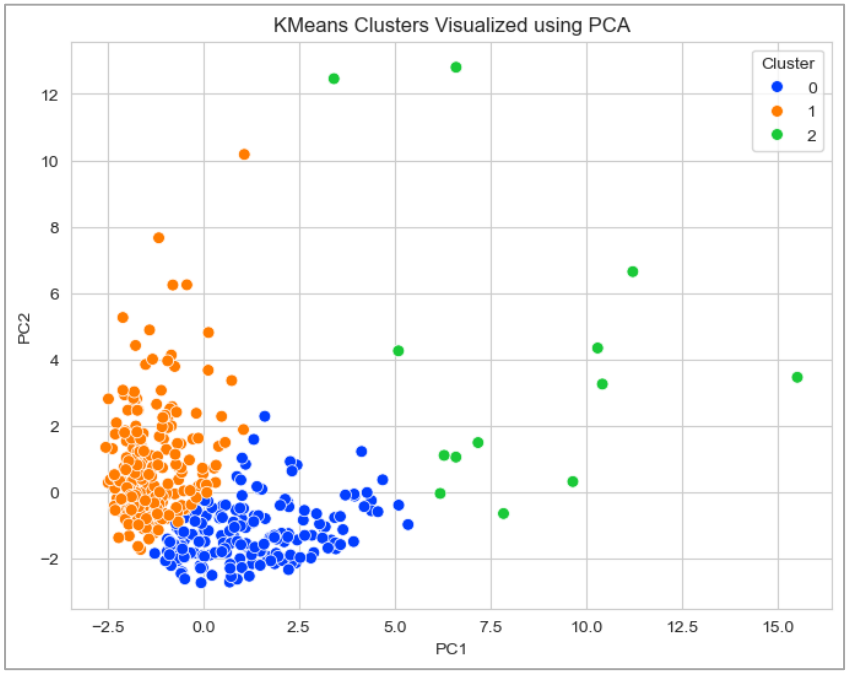
KMeans with k = 3 and random_state = 42 was trained on the standardized feature matrix. Each observation was assigned to one of the three clusters based on Euclidean distance to cluster centroids. The resulting cluster labels were appended to the dataset for further interpretation.

2.4 PCA-Based Cluster Diagnostics

To visually diagnose cluster quality, Principal Component Analysis (PCA) was performed on the same scaled dataset. The first two principal components captured most of the variance and clearly separated the clusters in 2D space:

- Cluster 0 formed a compact region with low perishable spending.
- Cluster 1 occupied a distinct region driven by Fresh and Frozen purchases.
- Cluster 2 appeared far along PC1, representing extreme high-volume buyers.

The PCA scatterplot confirmed the geometric separability of the KMeans clusters.



PCA Scatterplot showing clusters on basis of Channel

2.5 Validation and Model Diagnostics

Cluster quality was assessed through:

- Silhouette Score (k = 3): A positive score indicating meaningful separation.
- Cluster Size Distribution: Balanced segments (183, 244, and 13 customers), where the smallest cluster contains important high-spending clients.
- Cluster Profile Tables: Group-level means showed statistically distinct purchasing behaviours.

Together, these diagnostics demonstrate that the model is stable, interpretable, and supported by both statistical and visual evidence.

Part 3

Interpretation, Insights, Limitations, and Recommendations

3.1 Interpretation of Final Clusters

The clustering analysis successfully identified three distinct customer segments based on their purchasing behaviour across six product categories and engineered spending ratios. PCA visualizations showed clear separation between the clusters, confirming that the transformed and standardized features captured meaningful variation in purchasing patterns.

Cluster 0 — Small to Medium Retail Buyers

This segment shows high spending on Grocery, Milk, and Detergents_Paper, with very low spending on perishable items. These customers resemble convenience stores, mini-markets, and small retail outlets. Their purchases are regular, predictable, and dominated by packaged FMCG products.

Cluster 1 — HoReCa (Hotel–Restaurant–Café) Buyers

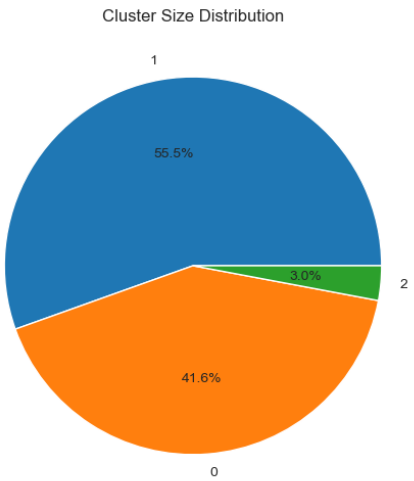
Customers in this segment exhibit high expenditure on Fresh, Frozen, and Delicassen products. This aligns with restaurants, hotels, catering services, and other food-service businesses. Their demand is perishable-heavy, less stable, and sensitive to seasonality and menu cycles.

Cluster 2 — High-Value Wholesale/Retail Chains

This is a small but extremely important segment. These customers spend heavily across all product categories, particularly Grocery and Detergents. They are likely hypermarkets, wholesalers, or large retail chains, contributing disproportionately to total revenue.

Cluster	0	1	2
Channel	1.584699	1.098361	1.846154
Region	2.579235	2.512295	2.615385
Fresh	4602.191257	16815.204918	25770.769231
Milk	7670.601093	2826.032787	35160.384615
Grocery	11250.306011	3664.139344	41977.384615
Frozen	1386.551913	4134.967213	6844.538462
Detergents_Paper	4584.775956	699.045082	19867.384615
Delicassen	1453.535519	1239.762295	7880.307692
Total_Spend	30947.961749	29379.151639	137500.769231
Perishable_Spend	5988.743169	20950.172131	32615.307692
Perishable%	0.205285	0.702005	0.206868
Retail_Spend	23505.683060	7189.217213	97005.153846
Retail%	0.740943	0.254975	0.739632

Cluster Profile



Cluster Distribution

These patterns reflect real-world customer heterogeneity and show that KMeans successfully captured three meaningful segments that differ both in intensity and type of product consumption.

3.2 Key Insights

- Clear Purchasing Archetypes
  The clusters correspond to well-defined commercial behaviours:
  - Retail/FMCG-oriented Buyers (Cluster 0)
  - HoReCa Perishable-Heavy Buyers (Cluster 1)
  - Bulk Wholesale Buyers (Cluster 2)

  This provides a strong foundation for targeted marketing and supply chain planning.

- Product Category Dependencies
  PCA loadings showed:
  - Grocery, Milk, and Detergents_Paper strongly influence PC1 (retail/FMCG dimension).
  - Fresh, Frozen, and Delicassen dominate PC2 (perishable dimension).

  This confirms two natural demand axes: retail vs. perishables.
  PCA confirmed two major behavioural dimensions: Retail Spending and Perishable Spending.

- High Value Concentration in Cluster 2
  Even though Cluster 2 contains only ~3% of customers, its spending levels are substantially higher across all features.
  This implies revenue concentration — a small customer segment drives a disproportionately large share of business that accounts for the highest overall consumption, making them a priority segment.

- Engineered Features Validate Interpretation
  Features like:
  - Perishable%
  - Retail%
  - Fresh–Frozen Ratio

  reinforced the segmentation by clarifying which customers prefer perishable vs. non-perishable categories.

- Channel-based validation demonstrated strong alignment: Cluster 1 is mostly Channel 1 (HoReCa), and Cluster 2 is mostly Channel 2 (Retail).

3.3 Limitations

- The dataset lacks demographic details such as customer size, revenue, or geography, which could enhance segmentation.
- Only one year of spending is available; behaviour may vary seasonally or annually.
- Clustering results may change with different scalers, feature sets, or distance metrics.

- A very small high-value cluster (Cluster 2) may require additional validation on larger data.

3.4 Recommendations for Business Use

➢ Targeted Marketing Strategies

- *Cluster 0 (Retail-Focused)*
  Promote bundled FMCG deals, loyalty discounts, or cross-category offers.
- *Cluster 1 (HoReCa)*
  Offer bulk pricing on perishables, customized delivery schedules, and subscription-based replenishment.
- *Cluster 2 (Bulk Buyers)*
  Dedicated account managers, volume contracts, priority inventory allocation, provide premium bulk discounts and exclusive deals

➢ Product Stocking & Inventory Strategy

- Increase perishable stock for Cluster 1-dominant regions.
- Strengthen supply chain for detergents, grocery, and milk in Cluster 0 zones.
- Maintain buffer inventory for Cluster 2 due to high variability and bulk orders.

➢ Customer Relationship Strategy:

- Assign dedicated account managers for Cluster 2 customers due to their high revenue impact.
- Provide loyalty incentives to stable FMCG buyers in Cluster 0.

➢ Revenue Optimization

i. Optimize pricing tiers for high-value customers (Cluster 2).
ii. Develop higher-margin product bundles targeted to Cluster 1 (e.g., meat + frozen + bakery items).

❖ Future Analysis Recommendations

1. Incorporate time-period data to analyze seasonal behaviour.
2. Add customer metadata (business type, geography, frequency) for richer segmentation.