Clustering Analysis of Wholesale Customer Data

Ismat Ara Khan

Abstract

This study applies clustering techniques to the Wholesale Customers dataset from Kaggle to segment customers based on purchasing patterns. Exploratory data analysis revealed correlations among product categories and highlighted variability in spending behavior. After preprocessing with StandardScaler and OneHotEncoder, K-Means and Hierarchical clustering were applied. The Elbow Method, Silhouette Score, and Dunn Index were used to determine the optimal number of clusters. K-Means with k=3 provided well-separated clusters, while hierarchical clustering preserved nested relationships. The analysis demonstrates that clustering can effectively identify customer segments, supporting targeted marketing and inventory planning.

1 Introduction

Customer segmentation is crucial in retail and wholesale businesses for designing targeted marketing strategies and managing inventory efficiently. This study explores unsupervised learning methods to identify distinct customer groups based on purchasing behavior. By analyzing spending patterns across product categories, clustering methods can reveal actionable insights into customer preferences and purchasing habits.

2 Data Overview

The dataset used in this study is the *Wholesale Customers Data Set* obtained from Kaggle [1]. It contains 440 records and 8 attributes, which include 2 categorical features (`Channel`, `Region`) and 6 numerical features representing annual spending in different product categories (`Fresh`, `Milk`, `Grocery`, `Frozen`, `Detergents_Paper`, and `Delicassen`). All columns have complete data without any missing values.

Exploratory Data Analysis (EDA) was conducted to examine the distribution of numerical and categorical features, detect potential outliers, and explore pairwise correlations among variables. This analysis helped in understanding patterns in the data and guided the preprocessing steps, including standardization of numerical features and encoding of categorical variables for clustering analysis.

3 Methodology

The following steps were performed to cluster the wholesale customer data [1]:

3.1 Data Preprocessing

Numerical features were standardized using *StandardScaler*, and categorical features (`Channel` and `Region`) were one-hot encoded to make them suitable for clustering algorithms [2].

3.2 Clustering Models

- **K-Means Clustering:** The K-Means algorithm [3] was applied to the preprocessed data. The optimal number of clusters was determined using the *Elbow Method* [4], which evaluates the within-cluster sum of squares (WCSS) for different k values.
- **Hierarchical Clustering:** Agglomerative Hierarchical Clustering [5, 6] was performed to explore the nested cluster structure and visualize the results using dendograms.

3.3 Cluster Validation

- *Silhouette Score:* Measures how similar an object is to its own cluster compared to other clusters. Higher values indicate better-defined clusters [7].
- *Dunn Index:* Evaluates cluster compactness and separation. A higher Dunn Index indicates well-separated and dense clusters [8].

3.4 Visualization and Comparison

Principal Component Analysis (PCA) was used to reduce the data to two dimensions for visualizing cluster separation [9]. The clustering results from K-Means and Hierarchical algorithms were compared using the Silhouette Score and Dunn Index to assess which method produced better-defined clusters.

4 Results and Discussion

Clustering analysis was performed using both K-Means and Hierarchical clustering algorithms. The evaluation metrics, including Silhouette Score and Dunn Index, were calculated for different numbers of clusters (k ranging from 2 to 9) and are summarized in Table 1.

Table 1: Comparison of Silhouette and Dunn Index for K-Means and Hierarchical Clustering

k	**Silhouette**		**Dunn Index**	
	K-Means	Hierarchical	K-Means	Hierarchical
2	0.3167	0.7138	0.0171	0.0690
3	0.4102	0.2578	0.0171	0.0162
4	0.3321	0.2620	0.0111	0.0260
5	0.3097	0.2876	0.0128	0.0270
6	0.2877	0.2938	0.0149	0.0318
7	0.2591	0.2022	0.0146	0.0318
8	0.2672	0.2052	0.0225	0.0318
9	0.2255	0.2126	0.0212	0.0318

- **K-Means:** The Silhouette Score was highest at $k = 3$ (0.4102), indicating moderately well-separated clusters, while the Dunn Index was relatively low across all k values, suggesting some overlap between clusters. PCA visualization showed clear separation among the clusters, confirming the clustering results showed in figure 1.
- **Hierarchical Clustering:** The Silhouette Score was highest at $k = 2$ (0.7138), indicating well-separated clusters for this configuration. The Dunn Index was also highest for $k = 2$ (0.0690), confirming that this clustering solution had more compact and well-separated clusters compared to higher k values. The dendrogram suggested 3 clusters, consistent with the moderate Silhouette Score observed at $k = 3$.

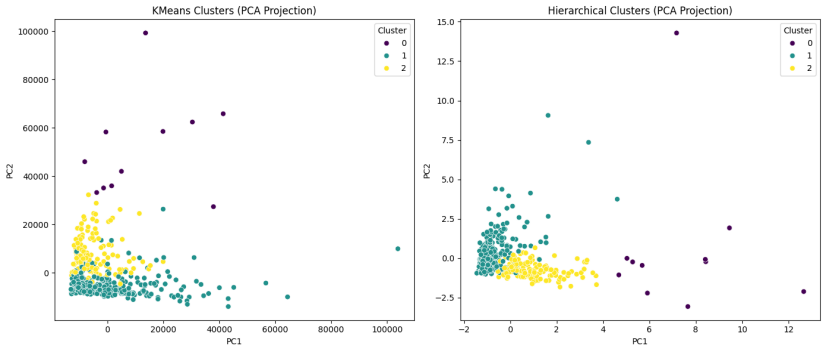

Figure 1: PCA projection of clusters for K-Means showing clear separation among clusters.

- **Comparison:** Comparing the two methods, Hierarchical clustering produced better-defined clusters for small k, as indicated by higher Silhouette Scores and Dunn Index values. K-Means achieved better scores at moderate k values but generally showed lower Dunn Index values than Hierarchical clustering, indicating less separation between clusters.

Overall, the results suggest that the choice of clustering algorithm and the number of clusters significantly affect cluster quality. Hierarchical clustering is more suitable when a smaller number of well-separated clusters is desired, while K-Means provides flexibility for moderate numbers of clusters but may result in slightly overlapping clusters. The PCA plots and dendrogram analysis corroborate these findings and provide visual confirmation of the cluster structure. These observations are consistent with the metrics presented in Table 1.

5 Conclusion

Clustering of wholesale customer data revealed three distinct customer segments based on purchasing patterns. K-Means produced compact, well-separated clusters, while hierarchical clustering preserved hierarchical relationships. Preprocessing and proper evaluation ensured meaningful clustering results. These insights can inform targeted marketing strategies and inventory management, demonstrating the practical value of unsupervised learning in customer segmentation.

References

[1] U. M. L. Repository, "Wholesale customers data set." https://www.kaggle.com/datasets/uciml/wholesale-customers, 2023.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[3] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, 1967.

[4] "Elbow method for determining the optimal number of clusters." https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set#The_Elbow_Method, 2023.

[5] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990.

[6] "Hierarchical clustering and dendrograms." https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html, 2023.

[7] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[8] J. Dunn, "Well-separated clusters and optimal fuzzy partitions." https://link.springer.com/article/10.1007/BF02291575, 1974.

[9] I. T. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Springer, 2nd ed., 2016.