Information Systems and Economics (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/infse.2024.050201
ISSN 2523-6407 Vol. 5 Num. 2

# *Research on Vegetable Replenishment and Pricing Strategy of Fresh Produce Superstores Based on K-means Cluster Analysis*

**Ye Tian[*], Sirong Cao, Xunxin Hu**

*Shenzhen University, Shenzhen, 518060, China*
*[*]Corresponding author: 2021190021@email.szu.edu.cn*

***Keywords:*** Spearman correlation analysis, K-Means cluster analysis, Simple exponential smoothing model

***Abstract:*** This study addresses the challenges of replenishment and pricing strategies for vegetable categories in supermarkets, considering the impact of a short shelf life on product quality. Employing Spearman correlation analysis, the paper investigate the relationships and distribution patterns among the sales volumes of different vegetable categories. With the overarching objective of profit maximization and focusing on the interplay between total sales volume and cost-plus pricing for various categories, we leverage theoretical frameworks such as K-Means clustering analysis and a simple exponential smoothing model. The resulting model is solved using tools like MATLAB, SPSSPRO, RStudio, etc. The derived replenishment and pricing strategies for each vegetable category offer substantial implications for enhancing supermarket operating profits.

## 1. Introduction

In fresh supermarkets, perishable vegetables generally have a short shelf life, and their quality deteriorates with increasing sales time, affecting consumer purchasing intentions and causing losses to the supermarket, resulting in a decline in benefits. In addition, supermarkets face space constraints in their sales area. Therefore, supermarkets typically replenish and price their products daily based on the historical sales and demand for each product.

Currently, there has been considerable research on replenishment and pricing of fresh products. Jiang Yingmei [1] and Mou Jinjin made joint decisions on inventory and pricing of fresh processed products based on perceived freshness. Li Yuan [2], considering the reference price effect, established joint optimization theories for pricing and inventory of single-channel retailers under dual replenishment modes. Qiao Xue [3] studied joint ordering and pricing strategies to mitigate the adverse effects of product decay and uncertain demand in the retail process of fresh products, addressing both quantity and quality losses. Wang Qiang [4] explored joint inventory and pricing strategies for retailers when consumers are risk-averse, and then used dynamic programming to study the retailer's multi-period optimal strategy by transforming the model equivalently. Hu Xinxue [5], considering uncertain demand, used fuzzy theory, utility theory, and mathematical methods to study ordering and pricing for organic agricultural product retailers. Kang Sha [6], based on different demand characteristics, constructed pricing and replenishment models for fresh agricultural products catering to homogeneous and heterogeneous demands. Chen Jun [7] and Kang Sha, aiming for profit maximization, used the perishable inventory theory to build a joint decision-making model for pricing and inventory replenishment for retailers in dual channels where demand depends on price and inventory levels, analyzing the nature of the existence of optimal solutions.

From the above literature, research on fresh products such as vegetables has focused on the reference price effect and optimal pricing and inventory decisions for retailers under dual-channel sales. In contrast, this paper, starting from the actual situation, compiles the sales volume of six major vegetable categories, analyzes the distribution patterns and relationships between different category sales volumes, and combines the commonly used cost-plus pricing method for vegetables. It constructs a replenishment and pricing strategy model for vegetable categories, which has significant economic implications (the data for this paper is sourced from http://www.mcm.edu.cn).

## 2. Analysis of Correlation between Individual Products, Product Categories, and Sales Volume

This chapter primarily conducts Spearman correlation analysis [8] on the weekly sales volumes between different categories and individual products, obtaining their relationships. Additionally, descriptive statistical analysis is performed on the weekly sales volumes for different categories and individual products to understand their distribution patterns. To facilitate model development, three assumptions are made here.

(1) It is assumed that different categories of vegetables are procured daily, meaning there are no categories without procurement on a given day.

(2) It is assumed that any remaining vegetables on a given day will not be sold the next day, implying there are no inventory issues.

(3) It is assumed that the sales price and volume of vegetables in the supermarket are unrelated to factors outside the data provided in the attachment.

### 2.1 Establishment and Solution of Spearman Correlation Analysis

(1) Calculate Spearman correlation coefficient

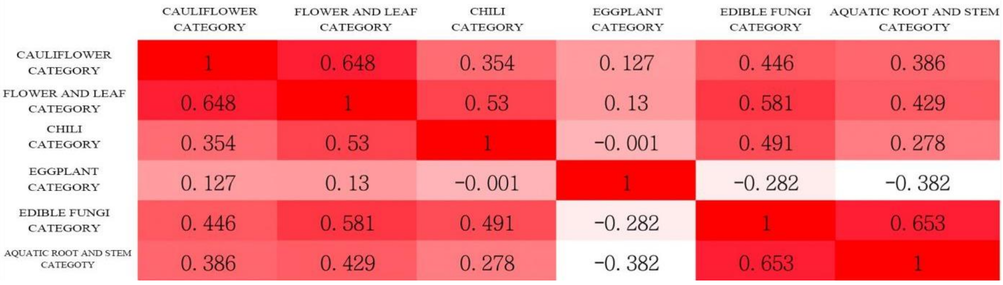
The formula for calculating the Spearman correlation coefficient is provided as follows:

$$\rho = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)})\cdot(R(y_i)-\overline{R(y)}\right)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)}\right)^2\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(y_i)-\overline{R(y)}\right)^2\right)}} \tag{1}$$

Especially, $x$ and $y$ denote different product categories or individual products, $R(x)$ and $R(y)$ are their respective ranks, and $\overline{R(x)}$, $\overline{R(y)}$ stand for the average ranks of $x$ and $y$ .The analysis involves computing these values to derive the Spearman correlation coefficient, providing insights into the relationships between different product categories and individual products.

(2) Solution

The analysis involves calculating the weekly sales volumes for different product categories or individual products. The obtained results reveal the correlation and significance levels between various product categories or individual products. Taking product categories as an example, the results are illustrated in Fig 1.



| | CAULIFLOWER CATEGORY | FLOWER AND LEAF CATEGORY | CHILI CATEGORY | EGGPLANT CATEGORY | EDIBLE FUNGI CATEGORY | AQUATIC ROOT AND STEM CATEGOTY |
|---|---|---|---|---|---|---|
| CAULIFLOWER CATEGORY | 1 | 0. 648 | 0. 354 | 0. 127 | 0. 446 | 0. 386 |
| FLOWER AND LEAF CATEGORY | 0. 648 | 1 | 0. 53 | 0. 13 | 0. 581 | 0. 429 |
| CHILI CATEGORY | 0. 354 | 0. 53 | 1 | -0. 001 | 0. 491 | 0. 278 |
| EGGPLANT CATEGORY | 0. 127 | 0. 13 | -0. 001 | 1 | -0. 282 | -0. 382 |
| EDIBLE FUNGI CATEGORY | 0. 446 | 0. 581 | 0. 491 | -0. 282 | 1 | 0. 653 |
| AQUATIC ROOT AND STEM CATEGOTY | 0. 386 | 0. 429 | 0. 278 | -0. 382 | 0. 653 | 1 |

Figure 1: Category Heatmap

## 2.2 Analysis of Model Results and Descriptive Statistical Analysis

Analysis of the correlation results for different categories reveals that most categories exhibit some degree of correlation. However, the sales data for Solanaceae (Tomato) category lacks significance with Cauliflower, Leafy Vegetable, and Chili categories, and shows very low correlation with Edible Fungi and Aquatic Root and Stem categories, with correlations of -0.282 and -0.382, respectively, suggesting no apparent correlation.

Among these, the most significant correlations are observed between Leafy Vegetable and Cauliflower categories and between Edible Fungi and Aquatic Root and Stem categories, with coefficients of 0.684 and 0.653, respectively, indicating a positive correlation. Their distribution is depicted in Fig 2. The curves of weekly sales volumes over time for each category demonstrate distinct periodic patterns.

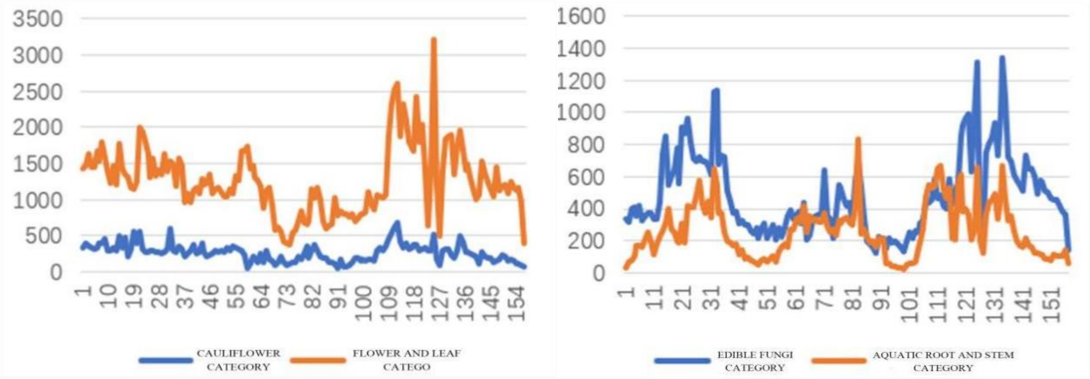


Figure 2: Comparison Chart of Weekly Sales Volume for Different Categories

There is a wide variety of individual products, totaling 246, and the correlation between the sales volumes of most categories is not significant. Additionally, the sales of individual products are influenced by supermarket procurement policies, consumer purchasing intentions, and agricultural cultivation conditions. Many individual products face suboptimal sales scenarios, resulting in numerous instances of zero values, which in turn contribute to correlation coefficients approaching 1. Therefore, in the analysis of the correlation between individual products, this paper focuses on analyzing products with good sales performance and correlation coefficients greater than 0.7. Two examples are illustrated in Fig 3.

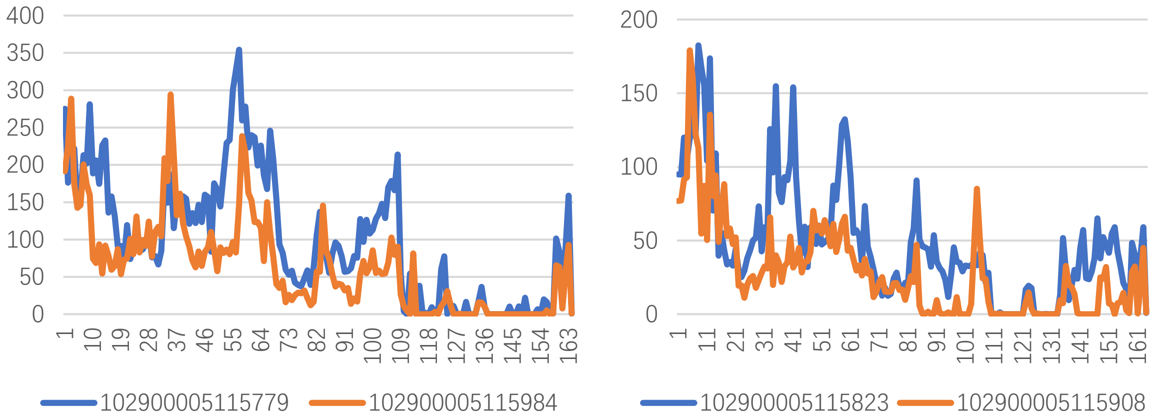


Figure 3: Comparison Chart of Weekly Sales Volume for Individual Products

In order to understand the distribution patterns of different product categories and individual products, this study employs descriptive statistical analysis on the weekly sales volumes, utilizing five key statistical measures: mean, standard deviation, quartiles, coefficient of variation, and the proportion of non-zero sales.

As shown in Fig 4, through the analysis of the weekly sales volume of six categories, it can be observed that in terms of weekly average sales, the Flower and Leaf category performs the best, reaching up to 1200 kilograms. Following that are the Chili and Edible Fungi categories, with sales ranging from 400 to 600 kilograms. Categories such as Eggplant and Cauliflower have relatively lower sales, below 300 kilograms. Additionally, influenced by the seasons, there is a notable variability in the weekly sales volume of categories, with certain periods experiencing a significant increase.

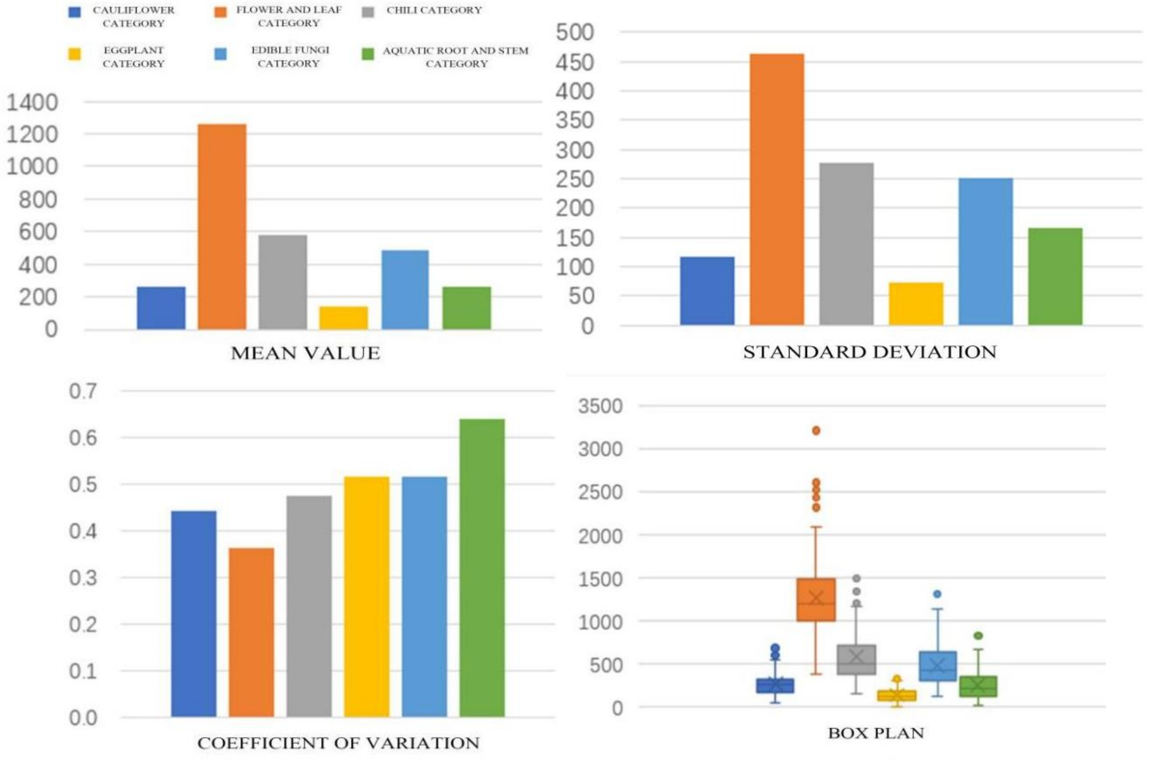


Figure 4: Distribution Pattern of Weekly Sales Volume by Category

The analysis of weekly sales volumes for individual products is illustrated in Fig 5. Firstly, based on the data of non-zero sales volume (= the number of weeks with sales volume > 0 / total number of weeks), it is observed that due to seasonality or buyer demand, most individual products experience sales or no sales during certain periods. It is noteworthy that 20 individual products, such as Fruit Chili (in orange), Kale Leaves, and Chinese Broccoli, had sales in only one week over the three years, while some individual products were sold for only a few weeks in the three-year period. Meanwhile, five individual products, including Broccoli, Clean Lotus Root (1), and Purple Eggplant (2), were available almost every week over the three years. Secondly, regarding the distribution of weekly sales volume for individual products, 63% of the individual products exhibit concentrated weekly sales volume data, indicating a strong seasonality for these products or a concentrated user demand during certain periods. Furthermore, looking at the average weekly sales volume, most fall within the [0, 18] range, with only a few individual products demonstrating outstanding weekly sales volumes, namely Broccoli, Clean Lotus Root (1), and Wuhu Green Pepper (1).

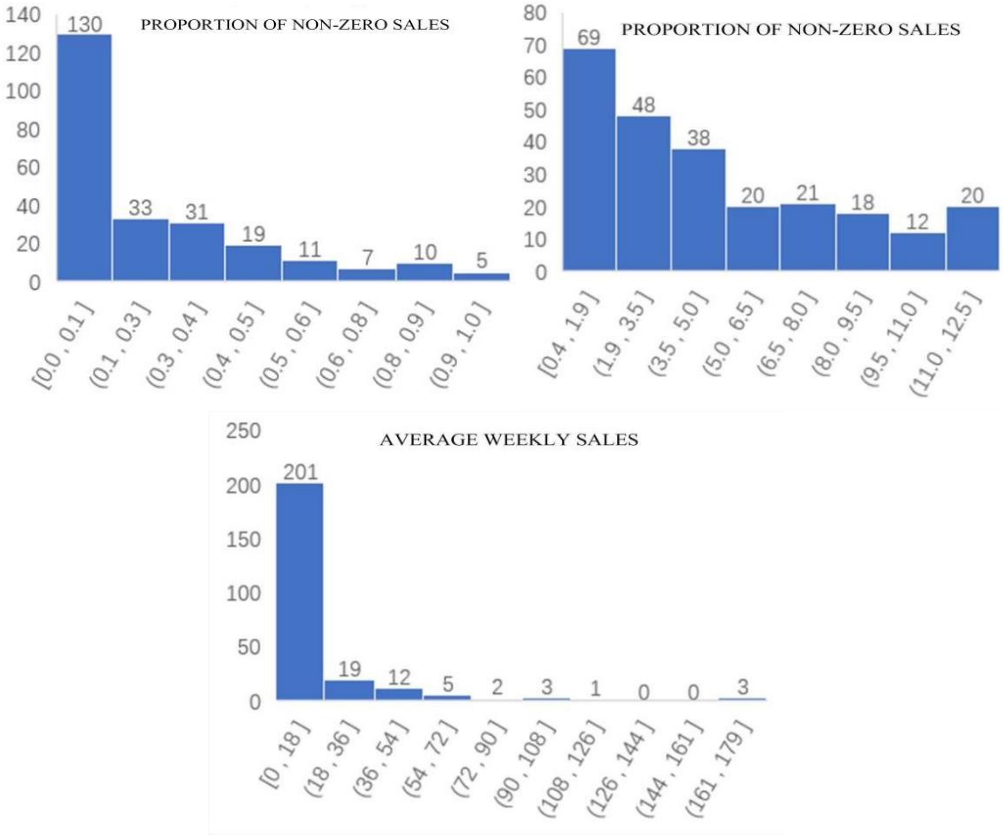


Figure 5: Summary Statistics of Weekly Sales Volume for Individual Products

## 3. Sales Volume and Cost-Plus Pricing: Formulating Marketing Strategies

In order to formulate replenishment and pricing strategies, it is necessary to explore the relationship between total sales volume and cost-plus pricing for different product categories. Based on this relationship, a replenishment and pricing strategy for the period from July 1, 2023, to July 7, 2023, will be developed to maximize the revenue for the supermarket.

To determine the mathematical model used to study the relationship between total sales volume and cost-plus pricing, this paper takes the flower and leaf category and the cauliflower category as examples. Initially, a scatter plot is created for the daily total sales volume and daily average cost-plus pricing of the categories. By observing the scatter plot in Fig 6 and conducting a thorough analysis, it is evident that there is no clear linear or non-linear relationship between the two variables. Therefore, the K-Means clustering method [9] is employed to explore the relationship as it yields the best results for these two variables.

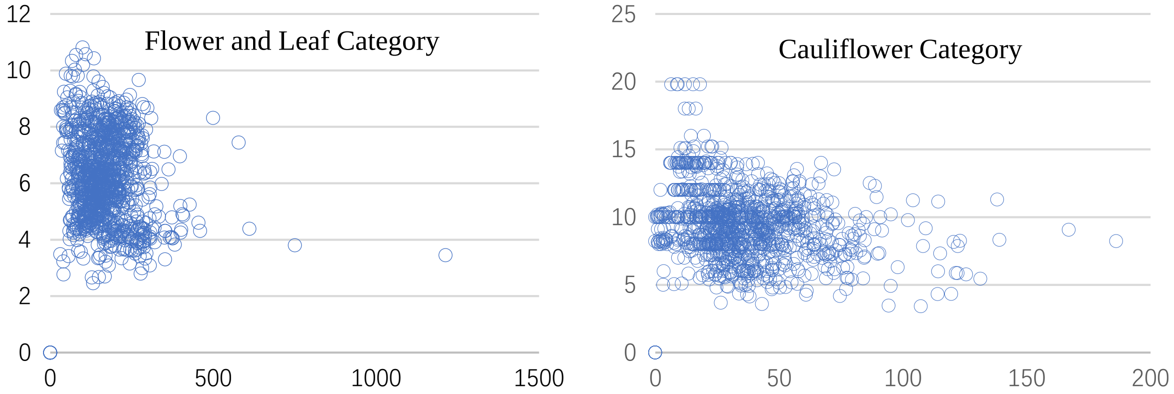


Figure 6: Scatter Plot of Category Sales Volume and Cost-Plus Pricing

Next, a line graph is generated for the daily total sales volume of different product categories, as shown in Fig 7. It can be observed that the sequences do not exhibit a clear trend. In traditional time series forecasting models, the fitting effect for such sequences is generally poor, with a goodness of fit often below 0.5. However, the simple exponential smoothing model [10] is effective in forecasting for sequences with strong volatility, making the fitting function stable. Therefore, this study utilizes the simple exponential smoothing model for forecasting. Since the forecast for the next 7 days falls under short-term forecasting, and the data itself is volatile and does not show a trend in the short term, it can be approximated that the changes in category sales volume within the next 7 days will generally remain within a small range.

To obtain the range of sales volume changes, this paper introduces a new indicator called "interval," calculated with the following formula:

$$\delta = \frac{|y_i' - y_i|}{n} \tag{2}$$

Where $y_i'$ is the predicted value, $y_i$ is the actual value, and $n$ is the number of forecast days.

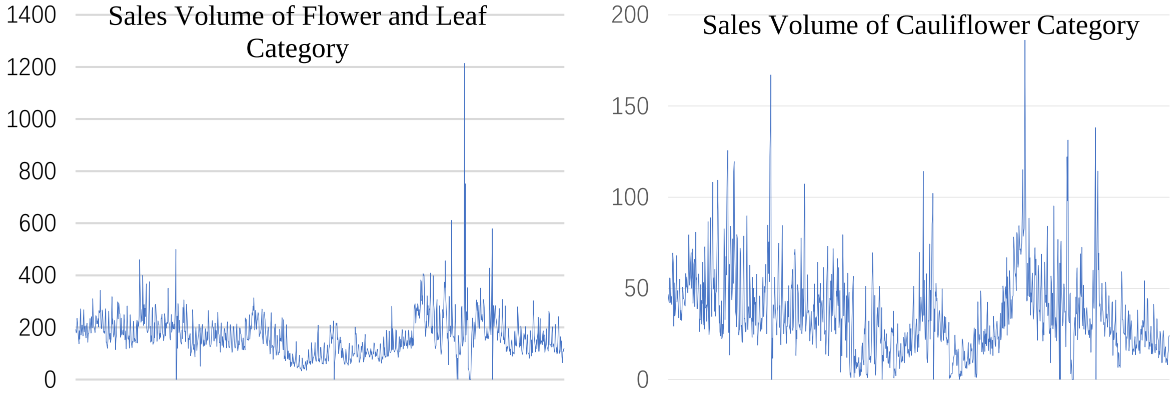


Figure 7: Line Chart of Daily Total Sales Volume by Category

## 3.1 Establishment and Solution of K-Means Clustering Analysis

(1) Determine the number of clusters k and the maximum iteration count n

In this study, k is set to 3, indicating that the daily total sales volume and daily average cost-plus pricing for different product categories will be clustered into three sets. To achieve optimal clustering results, the study does not set a specific iteration count but continues iterations until the convergence condition is met.

(2) Data Point Assignment

The program randomly selects three cluster centers. For each point in the dataset, it calculates the distance from each cluster center and assigns the point to the cluster where the distance is the smallest. The Euclidean distance formula is used for distance calculation.

$$d(x_i, x_j) = [\sum_{k=1}^{p} (x_{ik} - x_{jk})^2]^{1/2} \tag{3}$$

(3) Calculate the mean distance of the cluster and update the centroid

Recalculate the mean distance of each data point in the cluster to the sample center and update the centroid. Repeat steps 3 and 4 until there is no significant change in the data center points.

(4) Obtain the results

The K-Means clustering analysis for the six product categories is shown in Fig 8.

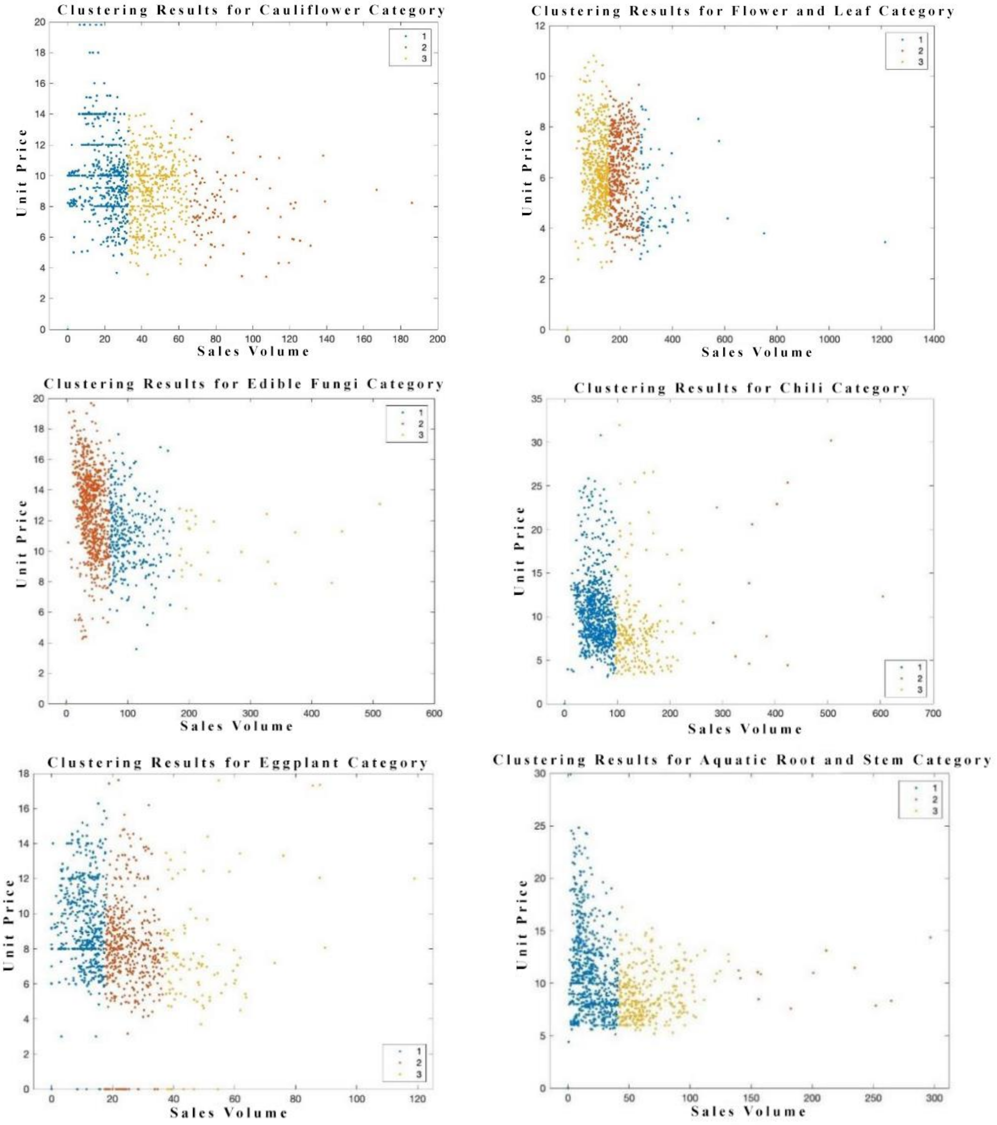


Figure 8: Clustering Analysis Results

The analysis of the model's strengths and weaknesses reveals that the average silhouette coefficient for different product categories reaches 0.58, DBI is consistently below 0.7, and the average CH is 1525, indicating an overall good clustering effect.

Through K-Means clustering analysis, the relationship between daily total sales volume and daily average cost-plus pricing for product categories can be established. Generally, the sales volume and cost-plus pricing relationship for most vegetable categories can be simplified into three classes: concentrated, moderately concentrated, and dispersed. It is also observed that at lower sales volumes, cost-plus pricing tends to be more concentrated. As sales volume increases, cost-plus pricing gradually becomes more dispersed, and the data points decrease.

Based on these findings, supermarkets can use historical data to predict sales volume and derive historical cost-plus pricing ranges. This allows for estimating approximate revenue in the face of uncertain procurement prices.

### 3.2 Establishment and Solution of Simple Exponential Smoothing Model

(1) Set Initial Values

Typically, this paper sets the initial value as the first value in the time series, i.e., the daily total sales volume data on July 1, 2022.

(2) Set the smoothing parameter $\alpha$

In the simple exponential smoothing model, the rate at which weights decrease is controlled by the parameter $\alpha$ , which ranges from 0 to 1. Considering the characteristics of the time series for daily total sales volume and to meet practical forecasting needs, this paper sets $\alpha = 0.8$.

(3) Calculate the Equation

To facilitate the generalization of this model, the paper expresses the equation in the following component form.

Prediction Equation:

$$y_{t+h\,|\,t} = l_t \tag{4}$$

Smoothing Equation:

$$l_t = \alpha y_t + (1-\alpha) l_{t-1} \tag{5}$$

In this context, $l_t$ represents the level component (or smoothing value) at time $t$. Setting $h=1$ yields the fitted value, and setting $t=T$ provides the genuine prediction beyond the training data.

(4) Solution

The results obtained through software are as follows. Partial results are shown in Fig 9, depicting the data fitting situations for Edible Fungi and Flower and Leaf from left to right.

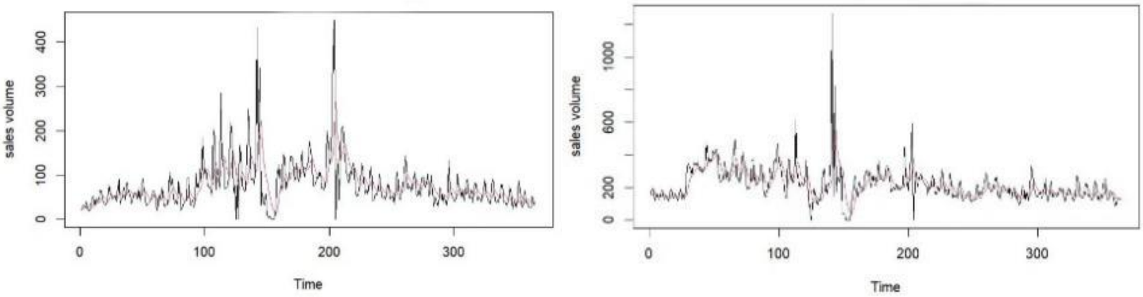


Figure 9: Simple Exponential Smoothing Model for Categories

The results for the total sales volume for the next 7 days are shown in Table 1.

Table 1: Summary of Seven-Day Sales

| Category Name | Predicted Value | $\delta$ | Prediction Interval |
|---|---|---|---|
| Cauliflower Category | 22.21 | 7.49 | [14.72,29.69] |
| Flower and Leaf Category | 130.49 | 41.1 | [89.39,171.6] |
| Chili Category | 81.38 | 29.03 | [52.35,110.41] |
| Eggplant Category | 19.56 | 4.45 | [15.11,24.02] |
| Edible Fungi Category | 37.57 | 22.22 | [15.35,59.79] |
| Aquatic Root and Stem Category | 11.85 | 12.78 | [0,24.62] |

Due to the characteristics of the exponential smoothing model, namely its relatively stable fitting function, the sales volume of product categories is not expected to undergo significant changes in the short term. Therefore, the table above provides only a single predicted value midpoint for each category, and the sales volume for each category in the next week is not expected to exceed the interval value around the predicted value.

By forecasting the total sales volume for the next 7 days and considering the loss rates for each category, the actual replenishment quantity for each category can be calculated. The predicted category replenishment quantity is calculated as follows: Predicted Category Replenishment Quantity = Predicted Category Total Sales Volume / Category Average Loss Rate. The results are presented in Table 2, which outlines the replenishment strategy for the next seven days.

Table 2: Replenishment Strategy

| Category Name | Average Loss Rate | Procurement Quantity Interval |
|---|---|---|
| Cauliflower Category | 15.51 | [17.42,35.14] |
| Flower and Leaf Category | 12.83 | [102.55,196.85] |
| Chili Category | 9.24 | [57.68,121.65] |
| Eggplant Category | 6.68 | [16.19,25.74] |
| Edible Fungi Category | 9.45 | [16.95,66.03] |
| Aquatic Root and Stem Category | 13.65 | [0,28.51] |

Based on the predicted total sales volume for the next seven days and considering the relationship between sales volume and cost-plus pricing, historical cost-plus pricing ranges can be provided. Supermarkets may consider pricing strategies based on potential factors such as customer loyalty, competition with other supermarkets, and overall profitability.

Considering the predicted total sales volume for the next seven days and the relationship between sales volume and cost-plus pricing, historical cost-plus pricing ranges can be established. Taking Fig 10, the relationship between sales volume and cost-plus pricing for the Flower and Leaf category, as an example, based on the predicted interval and the concentrated region of cost-plus pricing, the suggested cost-plus pricing range can be determined as [4.13, 8.88].

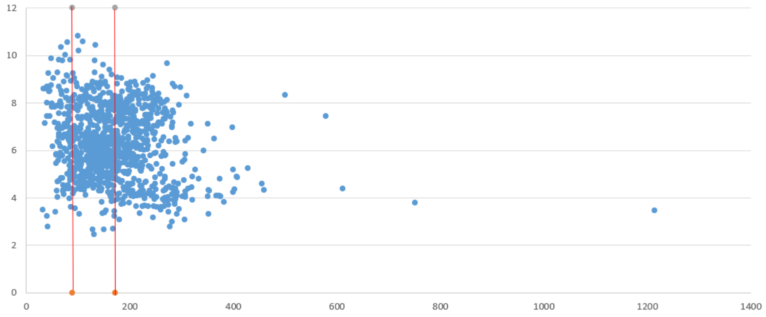


Figure 10: Relationship between Sales Volume and Cost-Plus Pricing for Flower and Leaf Category

Similarly, for the other six categories, the suggested cost-plus pricing ranges for the next seven days can be obtained as shown in Table 3.

Table 3: Suggested Cost-Plus Pricing

| Category | Cauliflower | Flower and Leaf | Eggplant | Edible Fungi | Chili | Aquatic Root and Stem |
|---|---|---|---|---|---|---|
| Price Range | [6.58,12] | [4,13,8.88] | [6.22,12.03] | [11.23,17.34] | [6.54,12.16] | [6.21,14.33] |

Therefore, the resulting profit ranges are shown in Table 4.

Table 4: Profit Range

| Category | Cauliflower | Flower and Leaf | Eggplant | Edible Fungi | Chili | Aquatic Root and Stem |
|---|---|---|---|---|---|---|
| Profit | [7.8016,176.6555] | [3.5756,821.964] | [12,2391,159.0122] | [88.416,709.7073] | [66.4845,760.7249] | [-35.4528,164.4616] |

Finally, the daily total profit range is determined to be [143.064, 2792.5255]. Research indicates that while increasing prices can enhance profit, supermarkets should consider potential factors such as customer loyalty and the pricing strategies of competing supermarkets. Historical data also shows negative cost-plus rates, suggesting that supermarkets should not solely pursue maximum profit but also consider other factors, even accepting certain losses for long-term benefits.

Therefore, this paper recommends supermarkets to adopt the median as a pricing strategy to balance the relationship between potential revenue and profit.

## 4. Conclusion

This paper, based on summarizing previous research, conducted a simple study on the issues of product sales and replenishment strategy from the following perspectives:

Firstly, from the perspective of categories and individual products, the relationship between products and sales volume was analyzed. Considering multiple angles can better explore the correlations between products. Supermarkets can strategically place highly correlated flower and leaf category and cauliflower category products in the same shopping area to attract customers. Additionally, based on the correlation between individual products, supermarkets can consider combining and promoting the sale of highly correlated individual products to stimulate consumption.

Secondly, considering that the scatter plots between daily sales volume and daily average cost-plus pricing are concentrated, this paper used K-Means clustering analysis to calculate the cost-plus pricing range for each category. This approach provides a more practical analysis of consumer buying behavior and serves as a theoretical basis for supermarket operators' procurement strategies, preventing losses due to unreasonable pricing.

Finally, through time series analysis to predict future sales volumes for each category and determine the replenishment quantity for each category, this paper established a simple exponential smoothing model to forecast sales volumes for the next seven days. Considering the relationship between sales volume and cost-plus pricing, it formulated a procurement strategy with the goal of maximizing revenue.

In summary, this paper, from multiple dimensions and building upon existing models, established a single-stage profit maximization model. However, the purchase price is not the sole factor affecting changes in sales volume. Factors such as product freshness and the supermarket's discount strategy can also alter consumer buying behavior, impacting supermarket revenue. Therefore, it is necessary to consider multiple factors comprehensively, addressing another focal point in the replenishment strategy problem. Additionally, the exponential smoothing model used in this paper can be further improved. For large-scale time series data and complex patterns, deep learning methods can provide more powerful predictive capabilities.

## References

*[1] Jiang Yingmei, Mu Jinjin. Joint Decision of Inventory and Pricing for Fresh Processed Products Based on Perceived Freshness [J]. Journal of Highway and Transportation Research and Development, 2020, 37(3): 151-158.*
*[2] Li Yuan. Research on Joint Pricing and Inventory Optimization of Dual-Channel Retailers Considering Reference Price Effect [D]. Yanshan University, 2023.*
*[3] Qiao Xue. Joint Replenishment and Pricing Strategy for Fresh Products Considering Sales Loss [D]. Southeast University, 2023.*
*[4] Wang Qiang. Research on Inventory and Pricing Strategy Considering Dual Replenishment Mode and Reference Price Effect [D]. University of Science and Technology of China, 2021.*
*[5] Hu Xinxue. Research on Ordering and Pricing Strategies of Organic Agricultural Products Retailers under Uncertain Demand [D]. Zhejiang University of Technology, 2020.*
*[6] Kang Sha. Research on Joint Pricing and Inventory Replenishment of Fresh Agricultural Products under Different Demand Characteristics [D]. Chongqing Jiaotong University, 2023.*
*[7] Chen Jun, Kang Sha. Joint Pricing and Inventory Replenishment Decision-Making for Agricultural Products with Dual-Channel Sales [J]. Industrial Engineering, 2023, 26(3): 39-46.*
*[8] Zhang Wenyao. Measuring Degree Correlation of Networks Using Spearman's Coefficient [D]. University of Science and Technology of China, 2016.*
*[9] Data Mining Center, Department of Statistics, Renmin University of China. Cluster Analysis in Data Mining [J]. Statistical Research, 2002(3): 4-10.*
*[10] Li Wenhong. Price Prediction of Eggs in Boxing County Based on Exponential Smoothing Model and ARIMA Model [D]. Shandong University of Finance and Economics, 2018.*