Research on Vegetable Product Replenishment Decision Based on K-means Algorithm

Yue Cao[1], Xingyue Yao[2], Mingqiao Yang[2], Xiunan Fang[1, *]

[1]College of science, JIAMUSI University, Jiamusi, China, 154007

[2]School of information and Electronic Technology, JIAMUSI University, Jiamusi, China, 154007

* Corresponding Author Email: fxntfx@sina.com

Abstract. The decision-making and pricing of supermarket restocking is a complex task that requires considering factors such as market demand, supply changes, and operating costs, making scientific and accurate decisions, providing customers with rich and fresh vegetable products, and promoting sustainable development. The pricing of vegetables in supermarkets generally adopts the "cost plus pricing" method, and supermarkets will formulate reasonable sales prices based on comprehensive considerations such as procurement costs, transportation costs, and operating costs. For products that may be damaged during transportation and have poor product quality, supermarkets usually offer discounts to meet customer needs and reduce food waste. This article first uses Excel software to describe and statistically analyze the data, transpose and visualize the data. Then, using the K-means algorithm in cluster analysis, the distribution patterns and interrelationships of vegetable categories and single product sales were analyzed in more detail. Finally, using Excel software, calculate the cost plus pricing for each vegetable category based on the formula "cost plus pricing=unit cost (1+cost profit margin)", and reasonable analysis is carried out according to the calculation results.This study on the calculation of cost-plus pricing for each vegetable category is of great significance to ensure the economic benefits of vegetable suppliers, improve the sustainability of agricultural production, optimize resource allocation, enhance market competitiveness, and promote the healthy development of the vegetable industry.

Keywords: Cost Plus Pricing, Clustering Analysis K-means Algorithm, Multi-Objective Programming Model

1. Introduction

From the demand side, there is a certain correlation between the sales volume and time of vegetable products in the replenishment and pricing decisions of supermarkets. Therefore, supermarkets can refer to historical sales data and market trends to predict the sales situation of the day and make replenishment decisions[1][2]. At the same time, supermarkets can also adjust their replenishment strategies based on market feedback and changes in customer demand, ensuring the freshness and variety of products[3]. This article focuses on analyzing the market demand for vegetable products to explore the corresponding problems[4].

From the perspective of the supply side, the variety of vegetables is relatively abundant in April and October. Supermarkets need to arrange their sales mix reasonably, and timely replenish according to factors such as the origin, shelf life, and sales volume of different vegetable varieties to ensure sufficient supply[5]. Supermarkets also need to consider the limitation of purchase transaction time between 3:00-4:00 in the morning, which may make it difficult to accurately understand specific individual products and purchase prices. Therefore, they need to rely on various market information and procurement channels to flexibly respond and ensure the quality of vegetable supply.

2. Research on the sales volume of various categories of vegetables

2.1. The structure of BP neural network

Vegetables in this article are divided into six categories: flowering and leafy vegetables, cauliflower, aquatic rhizomes, eggplants, chili peppers, and edible mushrooms. Among them,

flowering and leafy vegetables are divided into 101 single products, including beef head lettuce and Sichuan red Chinese toon; Cauliflower vegetables are divided into 5 types: broccoli, purple cabbage, etc; Aquatic rhizomes are divided into 19 types, including lotus root and clean lotus root; Eggplant vegetables are divided into 10 types: purple eggplant, green eggplant, etc; Chili vegetables are divided into 45 types of single products, including red chili peppers and green chili peppers; Edible mushrooms and vegetables are divided into 72 types, including Xixia mushroom and Xixia shiitake mushroom[6]. In order to better analyze the distribution pattern of sales volume of various categories of vegetables and individual items of vegetables, this article decides to analyze the average, mode, standard deviation, maximum, and minimum sales volume of each category of vegetables and individual items of vegetables by calculating the average, mode, standard deviation, maximum, and minimum values of sales volume of each category of vegetables[7]. Based on time information, vegetable categories are divided into four seasons to analyze the sales patterns of each category of vegetables in the four seasons. For example, in spring, whether the sales volume of cauliflower vegetables is higher, and whether the sales volume of chili vegetables is ahead of other categories of vegetables in winter; Divide each individual vegetable item according to the degree of packaging, and observe whether the price of the same item is the same under different packaging. The following is a detailed analysis of the relevant charts made in conjunction with Excel software as shown in Fig.1:

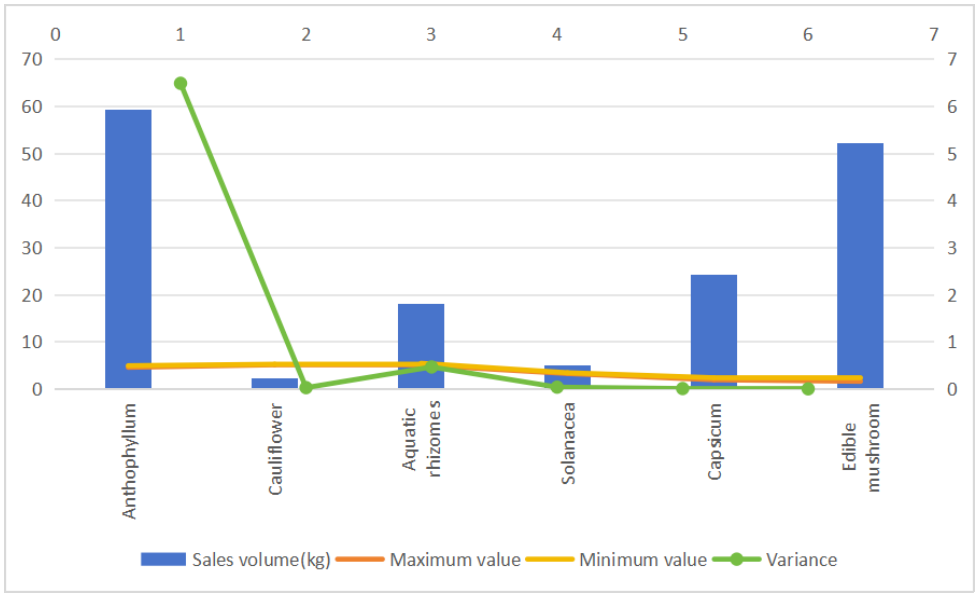

Fig. 1 Sales of various categories of vegetables

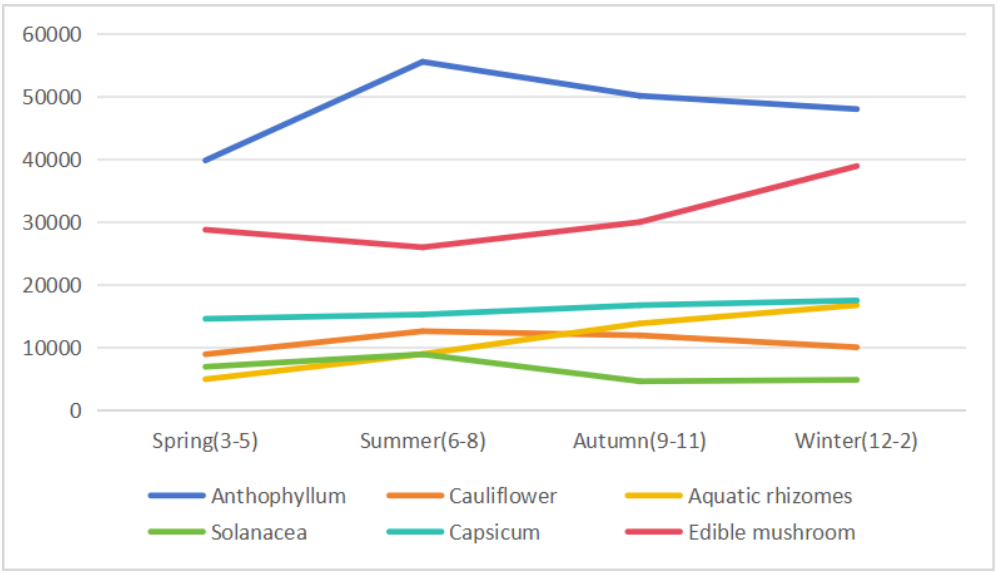

Fig. 2 Sales of various categories of vegetables in the four seasons

From Fig.2, it can be seen that the sales volume of flower and leaf vegetables is always the highest throughout the four seasons. From a demand side perspective, the average sales volume of flower and leaf vegetables is the highest among the six varieties, indicating a high demand for flower and leaf vegetables and a high supply of goods in supermarkets; Considering the time series of the four seasons, the average sales of each category are relatively stable, and their distribution pattern is relatively balanced.

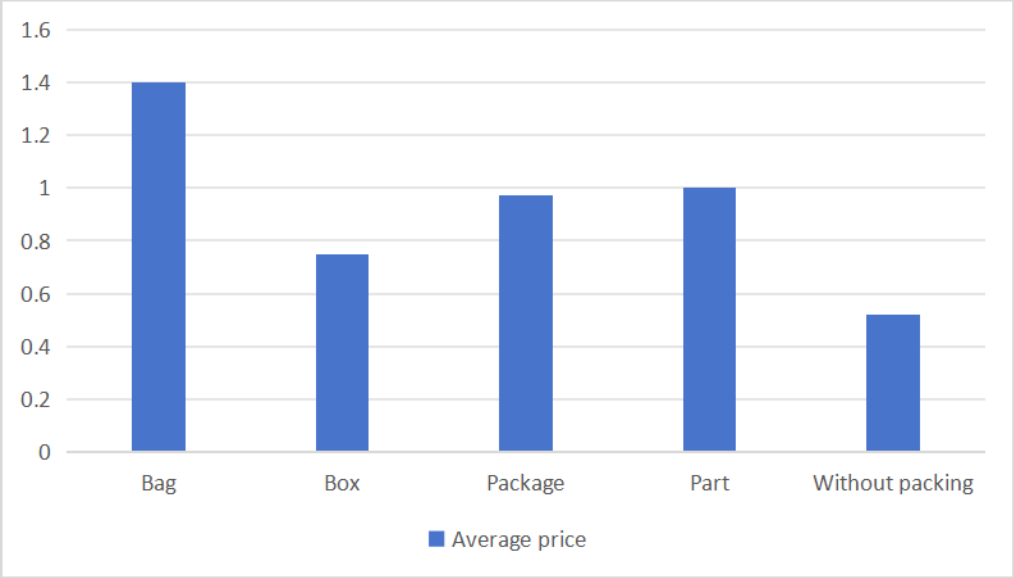

Fig. 3 Sales prices of different purchasing methods

From Fig.3, it can be seen that 251 individual products are classified according to the sales prices of different purchase methods. From the graph, it can be seen that the average purchase price of a single product in bagged form is much higher than the average price of other forms of packaging. The sales price of unpackaged form is the lowest, indicating that bagged vegetables have higher convenience in the retail process, and bagged vegetables usually undergo additional processing and screening to ensure product quality and freshness, resulting in increased costs. The price of bagged vegetables is relatively high.

2.2. Constructing a clustering analysis model and solving it

In order to further observe the interrelationships between different categories of vegetables and the sales volume of individual products, and because the data involved in the article is continuous data, this article has decided to use clustering analysis K-means algorithm to solve this problem. K-means algorithm is a measure of similarity between data objects using Euclidean distance, which is inversely proportional to the distance between data objects[8]. The larger the similarity, the smaller the distance. Therefore, it is selected to solve the interrelationships between the sales volume of various categories of vegetables and individual products, and determine whether there are significant differences or correlations between their categories and individual products.

(1) K-means algorithm in cluster analysis

Select K value

Firstly, randomly select K initial cluster centers $C_i(\mathrm{i} \leq 1 \leq \mathrm{K})$ from the corresponding data

Calculate Euclidean distance.

The formula for calculating Euclidean distance is:

$$\mathrm{d}(\mathrm{X}, \mathrm{C_i}) = \sqrt{\sum_{j=1}^{m}(X_j - C_{ij})^2} \tag{1}$$

Among them, X is the data object; C_i is the i-th cluster center; M is the dimension of the data object; X_j, C_{ij} is the jth attribute value of X and C_{ij}.

2.3. Specific classification of algorithm steps

(1) The silhouette coefficient silhouette value is used to measure the similarity of clustering, with a range of -1 to 1. The larger the value, the better the node matches its belonging cluster rather than adjacent clusters. Therefore, the closer the value is to 1, the better the clustering effect[9].

Here, three clusters are selected and combined in pairs, with a K value of 3 and a cluster center of C_1、C_2、C_3, as shown in Fig.4.

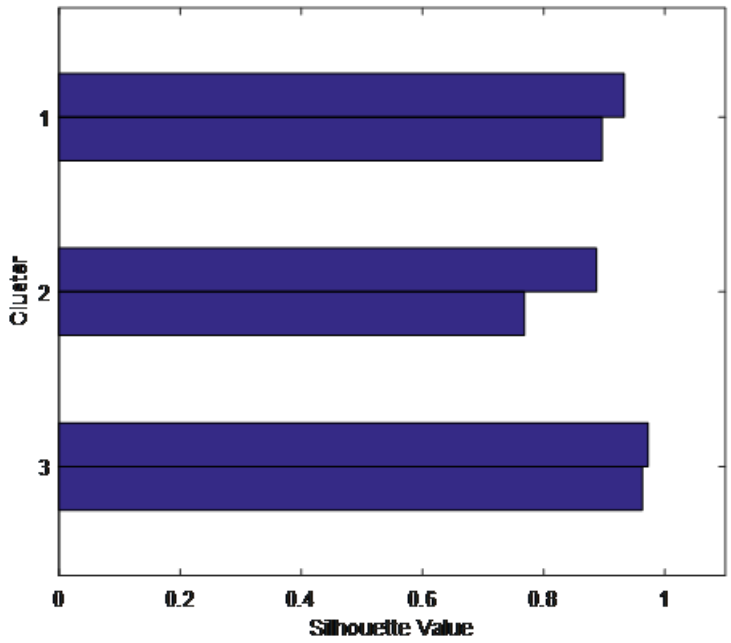

Fig. 4 Cluster analysis outline of various categories of vegetables

(2) Cluster scatter plot

Assign data objects to the cluster corresponding to the cluster center C_1、C_2、C_3 by calculating Euclidean distance.

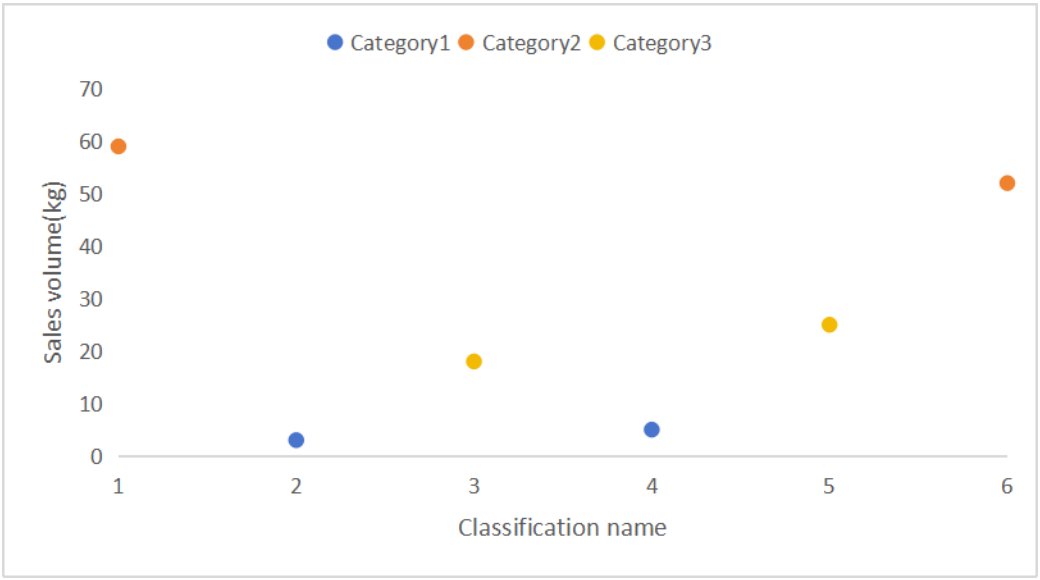

Fig. 5 Clustering scatter plot of various categories of vegetables

From the colors in Fig.5, it can be seen from the clustering analysis contour map of various vegetable categories that when K=3, the contour coefficient is closest to 1. Therefore, the paper classify the flowering and edible fungi into one group; Cauliflower and eggplant are classified into one category; Aquatic rhizomes and chili peppers are classified into one category. The correlation between categories grouped together is high.

2.4. Categorize vegetables by individual item

(1) Profile coefficient

Here, 6 clusters are selected, and the screened 245 single products are divided into 6 categories, with a K value of 6, which means the cluster centers are C_1、C_2、C_3、C_4、C_5、C_6.

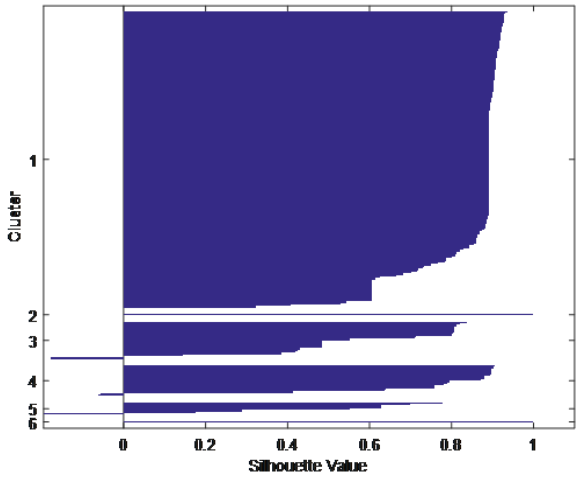

Fig. 6 Cluster analysis outline of each individual vegetable product

(2) Cluster scatter plot

By calculating the Euclidean distance, the data objects are assigned to the clusters corresponding to the cluster centers C_1、C_2、C_3、C_4、C_5、C_6 and.

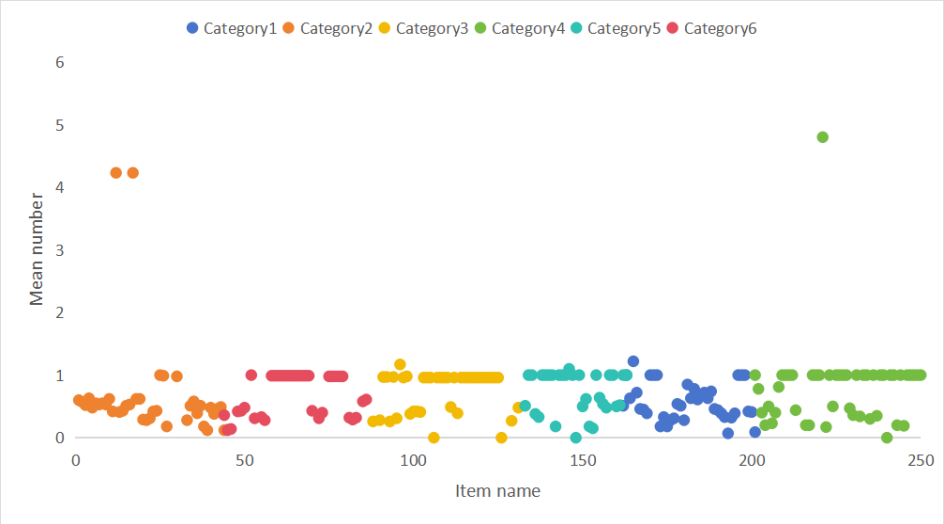

Fig. 7 Cluster scatter plot of various categories of vegetables

Based on the analysis of Figs 6 and 7, it can be seen that the last 245 items are divided into six categories and classified using six different colors. Items classified in the same category have similar average sales volumes, indicating that items classified in the same category have strong interrelationships.

3. One variable regression prediction model and its solution

It is known that vegetable categories are divided into six categories: flower and leaf, cauliflower, aquatic rhizomes, eggplants, chili peppers, and edible mushrooms[10]. Here, the paper first take edible mushroom vegetables as an example, and set the total sales volume of edible mushroom vegetables as the independent variable; Using the cost markup pricing of edible mushroom vegetables as the dependent variable, establish a univariate linear regression model, whose mathematical model can be written as:

$$y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i \tag{2}$$

Where β_0, β_1 is the regression coefficient and ε is the random error.

(1) Variable assumption:

The explanatory variable $x_1, x_2, \ldots, x_m$ is a non random variable, not a random variable, and the observed value $x_{i1}, x_{i2}, \ldots, x_{ip}$ is a constant.

$$\mathrm{E}(\varepsilon_\mathrm{i}) = 0, \mathrm{i} = 1,2,\ldots,\mathrm{m} \tag{3}$$

(2) Equivariance and irrelevant assumptions:

$$\mathrm{cov}(\varepsilon_\mathrm{i}, \varepsilon_\mathrm{j}) = \begin{cases} \sigma^2, i = j \\ 0, i \neq j \end{cases} \; i,j = 1,2,\ldots,\mathrm{m} \tag{4}$$

(3) Assumption of normal distribution:

$$\begin{cases} \varepsilon_i \sim N(0, \sigma^2), i = 1,2,\ldots,m \\ \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m \end{cases} \tag{5}$$

(4) The number of sample sizes n is greater than the number of explanatory variables p.

3.1. Least squares estimation of regression coefficients

According to the model assumption, taking into account factors such as the total sales volume of each vegetable category and the cost plus pricing of each vegetable category, solve for the partial derivatives of β_0 and β_1 and make them 0. After sorting, the equation system is:

$$\begin{cases} n\beta_0 + (\sum_{i=1}^n x_i)\beta_1 = \sum_{i=1}^n y_i \\ (\sum_{i=1}^n x_i)\beta_0 + (\sum_{i=1}^n x_i^2)\beta_1 = \sum_{i=1}^n x_i y_i \end{cases} \tag{6}$$

The linear regression equation can be obtained by solving the equation system as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 \tag{7}$$

3.2. Testing of Regression Equations and Regression Coefficients

(1) Find the sum of squared total deviations of y_i

$$\mathrm{SST} = \sum_{i=1}^n (y_i - \bar{y})^2 \qquad \text{degree of freedom } df_T = n - 1 \tag{8}$$

(2) Finding the sum of regression squares for y_i

$$\mathrm{SSR} = \sum_{i=1}^n (\hat{y} - \bar{y})^2 \qquad \text{degree of freedom } df_R = 1 \tag{9}$$

(3) Finding the residual sum of y_i

$$\mathrm{SSE} = \sum_{i=1}^n (y_i - \hat{y_i})^2 \qquad \text{degree of freedom } df_E = n - 2 \tag{10}$$

(4) Find the decomposition formula for the sum of squares

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE} \tag{11}$$

Goodness of fit test

$$R^2 = 1 - \frac{SSE}{SST} \approx 0.897 \tag{12}$$

Due to the goodness of fit R^2 value of 0.897 approaching 1, it can be seen that the regression equation has a good fitting effect at this time, indicating that further analysis can be carried out.

Construct an F-test statistic, where $\mathrm{F} = \frac{SSR}{SSE/(n-2)} \sim F(1, n-2)$ tests the significance of the regression equation and proposes null and alternative hypotheses for this:

H_0: Assuming that the sales volume of edible mushroom vegetables has no significant impact on the cost markup pricing of edible mushroom vegetables, H_1: Assuming that the sales volume of edible

mushroom vegetables has a significant impact on the cost markup pricing of edible mushroom vegetables.

(1) If $F > F_{0.01}(1, n-2)$, then the linear relationship is extremely significant;

(2) If $F_{0.05}(1, n-2) < F < F_{0.01}(1, n-2)$, then the linear relationship is significant;

(3) If $F < F_{0.05}(1, n-2)$, there is no linear relationship.

Combining the above algorithms, use SPSS to create an analysis of variance table, as shown in Table 1:

Table 1. Analysis of Variance

ANOVAa						
model		Sum of squares	degree of freedom	mean square	F	Significance
1	regression	2769.460	1	2769.460	96265.279	0.000b
	residual	675.037	23464	0.029		
	total	3444.497	23465			
a. Dependent variable: sales of edible mushrooms						
b. Predictive variables: (constant), mushroom cost markup pricing						

According to the chart of analysis of variance, it can be seen that the p-value is less than 0.05, so there is a significant relationship between the cost markup pricing of edible mushrooms and their sales volume.

Similarly, it can be inferred that there is no significant relationship between the cost markup pricing of cauliflower and the total sales volume of cauliflower; There is a significant relationship between the cost markup pricing of flowers and leaves and their sales volume; There is a significant relationship between the cost markup pricing of chili peppers and their sales volume; There is a significant relationship between the cost markup pricing of eggplants and their sales volume; There is a significant relationship between the cost markup pricing of aquatic rhizomes and their sales volume.

4. Conclusions

This article uses Excel software to describe and statistically analyze the data, transpose and visualize the data, and finally uses the K-means algorithm in cluster analysis to analyze the distribution patterns and interrelationships of vegetable categories and single product sales in more detail. Meanwhile, the model adopted in this article can select appropriate statistical methods for quantitative analysis based on the classification of data and the requirements of the problem, and the model is relatively easy to test. K-means clustering analysis converges to local optima and is suitable for clustering high-dimensional data. The use of multi-objective programming models can consider multiple factor constraints and multi-objective combinations, and ultimately achieve the optimal goal. The multiple linear regression prediction model has strong interpretability and high accuracy. Using the multiple linear regression prediction model to explore its correlation can quickly predict a large number of datasets, and the results are also easier to understand.

References

[1] Si Shoukui, Sun Xijing. Python Mathematical Modeling Algorithm and Application [M]. Beijing: National Defense Industry Press,2023.

[2] He Xiaoqun, Liu Wenqing. Applied regression analysis[M]. 5th Ed. Beijing: China Renmin University Press,2019.

[3] Sheng K ,Mei X ,Yang X , et al. Research on Pricing and Replenishment of Vegetable Products Based on the Optimization of Supermarket Revenue[J]. Information Systems and Economics,2024,5(1).

[4] Li Fangfang, Li Xiufang. Pricing decision model of property insurance company based on Multi-objective planning theory [J]. Journal of Zhongnan University of Economics and Law,2015(01):48-54.

[5] Long C . Optimization research on automatic pricing and replenishment decision method for vegetable commodities[J]. Advances in Engineering Innovation,2024,7(1).

[6] Fan F Y. Research on pricing Model of fresh commodities in supermarket [D]. Fujian Normal University,2020.

[7] Wang X ,Wang X . Optimizing the Sales Law and Replenishment Decision Analysis of Fresh Supermarket Products[J]. Frontiers in Computing and Intelligent Systems,2024,7(2).

[8] Yu J ,Li J ,Wang S . Research on the optimization of vegetable sales and pricing strategy of single products[J]. International Journal of New Developments in Engineering and Society,2023,7(9).

[9] Zhao J. Vegetable pricing and replenishment decision of supermarket based on optimal profit planning model [J]. The circulation economy, 2024 (9) : 8-12.

[10] Tian Y ,Cao S ,Hu X . Research on Vegetable Replenishment and Pricing Strategy of Fresh Produce Superstores Based on K-means Cluster Analysis[J]. Information Systems and Economics,2024,5(2).