A Study of Replenishment And Pricing Strategies For Vegetable Commodities Based On Multiple Linear Regression

Chujun Luo[1, #], Hongtao Min[2, #], Xiaohan Hu[1], Yining Wu[3, *]

[1]The financial collage, Jiangxi Normal University, Nanchang, China, 330022

[2]School of Computer Information Engineering, Jiangxi Normal University, Nanchang, China, 330022

[3]College of Finance and Statistics, Hunan University, Changsha, China, 410082

* Corresponding Author Email: Evelyn721@hnu.edu.cn

#These authors contributed equally.

Abstract. Vegetables, which can provide the human body with rich nutrients such as vitamins, minerals and dietary fiber, are indispensable necessities in people's daily life. Therefore, on the one hand, in order to ensure the quality of vegetables to provide fresh vegetables, superstores need to replenish goods in a timely manner; on the other hand, in order to improve the revenue, superstores need to reasonable pricing. This paper takes 251 kinds of vegetables sold in fresh food superstores as the research object, firstly, using Pearson's correlation coefficient to explore the correlation between the sales volume of vegetable single product and category, and obtains that the strongest correlation in vegetable category is cauliflower and edible mushrooms, with a correlation coefficient of -0.9989, and that the strongest correlation in vegetable single product is Hericium erinaceus and fresh zongzi leaves, with a correlation coefficient of 0.9973. Then, this paper set up a multivariate linear regression model to solve the problem for each vegetable. Linear regression model to solve the relationship between total sales volume and cost-plus pricing for each category. Finally, for the total daily replenishment, through the total daily replenishment = total predicted sales / (1 - wastage rate) processing to get the total daily replenishment of vegetables for 7 days, to obtain the cost matrix and markup rate matrix, and then get the pricing of vegetables for 7 days. For example, the replenishment of cauliflower on Monday should be 268.7238 kg, and the pricing is $2.5179/kg.

Keywords: Pearson Correlation Coefficient, Multivariate Linear Regression Model, Cost Bonus Pricing.

1. Introduction

As the saying goes, "Food is the God of the people." With the rapid development of the economy and the gradual enrichment of material life, people's requirements for living standards are getting higher and higher, and the concept of consumption of vegetables and other foods has also produced a huge change in the freshness of vegetables and other quality requirements gradually improved. From 2013 to 2021, the per capita vegetable consumption of Chinese residents will increase from 94.9kg to 106.2kg, with a growth rate of 11.9% and an average annual growth rate of 1.3% [1]. As an important part of the circulation system, fresh food supermarket chains and agricultural supermarkets to ensure the supply of vegetables and other important agricultural products on the market to highlight the important role of sales channels [2]. Vegetables are seasonal, perishable, intolerant of transportation and other characteristics, belonging to the perishable goods, easy due to collision damage, corruption, drying or gasification and other reasons in the storage time is prone to death or deterioration [3]. In addition, affected by temperature and sensitivity to time, vegetables need to consider the quality and safety issues in the circulation process [4]. Therefore, as the main channel for residents to purchase vegetables, superstores will replenish the stock every day according to the historical sales and demand of goods.

Due to the dispersed distribution of the origin of vegetables as well as climate and cost factors [5], coupled with the high demand of consumers for the quality and freshness of vegetables, superstores usually replenish the stock between 3:00 a.m. and 4:00 a.m. every day, and at the same time, they adopt the "cost-plus pricing" method for vegetables, and discount the damaged goods in order to improve the quality and safety of vegetables at the same time and at the same time, to ensure the quality of vegetables. Window constraints, total distribution costs, fixed investment costs, and handling costs to ensure the freshness of vegetables while minimizing replenishment costs and wastage costs [6]. To make rational replenishment and pricing decisions, reliable supply and demand relationships and market analysis are particularly important.

In this paper, the total sales volume of vegetables during the period from July 2020 to June 2023 is firstly accumulated and visualized, and histograms and three-dimensional pie charts are used to show the distribution of high and low sales volume and sales volume share of each category, and distribution line graphs are used to show the sales volume of 251 individual vegetable products, so as to analyze the distribution pattern of vegetable categories and sales volume of individual products. After that, Pearson correlation coefficient method is used to analyze the interrelationship between the sales volume of different categories or different single products of vegetables. Finally, a multiple linear regression model was constructed with the total sales volume of each vegetable category as the dependent variable and the corresponding sales price and wholesale price as the independent variables, and then Matlab was used to draw a three-dimensional stereoscopic image to describe the correlation in a clear and intuitive way. After that, we use the "cost-plus formula" to calculate the appropriate pricing, and give the total daily replenishment and pricing strategy of the vegetable category in the next week in order to realize the maximum profit.

2. Potential links between different categories of vegetables

First we preprocess the data, and then we establish the mapping relationship between the 5-digit item code number and the category in Matlab, so that we can quickly find the category corresponding to a single item of the given data through the relationship, and then we perform the statistics of the total amount of sales of these individual items by category.

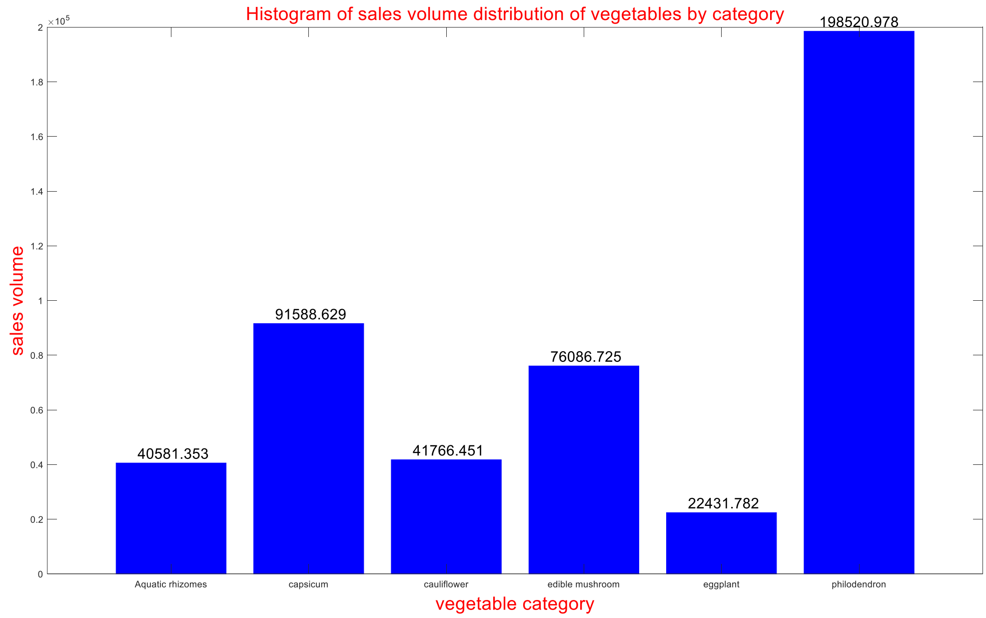

Figure 1. Histogram of sales volume distribution of vegetables by category

For the potential connection between different categories of vegetables we mainly through the establishment of charts to illustrate the distribution pattern and correlation between the sales volume of individual vegetable items and each category. Figure 1 shows the distribution law of vegetable categories by showing how much sales volume of each category of vegetables, we can see that the difference between the flower and leafy vegetables with the most sales volume and the eggplant vegetables with the least sales volume is $106{,}932.349\times10^{5}$kg higher than that of the chili vegetables with the second highest sales volume of $91{,}588.629\times10^{5}$kg, which is higher than the second highest sales volume of chili vegetables. It can be said that the dominant position of leafy vegetables is very obvious. This is mainly due to the fact that leafy vegetables have high attraction in taste, nutritional value and appearance, and at the same time, their planting cost is relatively low and their yield is relatively high.

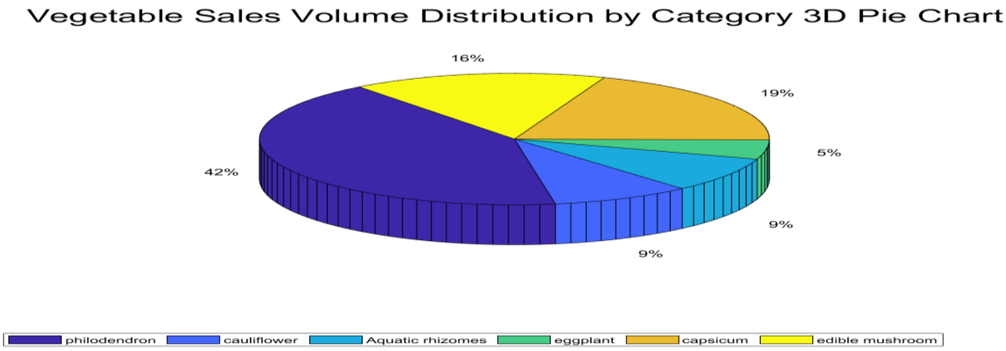

Figure 2. Vegetable Sales Volume Distribution by Category 3D Pie Chart

Figure 2 is a three-dimensional pie chart of the distribution of sales volume of vegetables by category, and it can be clearly seen that the sales volume of leafy and flowering vegetables is the highest and the sales volume of eggplant vegetables is the lowest among all categories. Based on the sales of leafy vegetables and eggplant vegetables in Figure 1, we can make the following two conjectures: first, leafy vegetables have the highest sales volume among all categories of vegetables, probably because consumers are more inclined to buy these vegetables, or they have higher nutritional value or delicious taste. On the contrary, eggplant vegetables have the least sales volume, which may indicate that consumers do not like their taste and their nutritional value is not high.

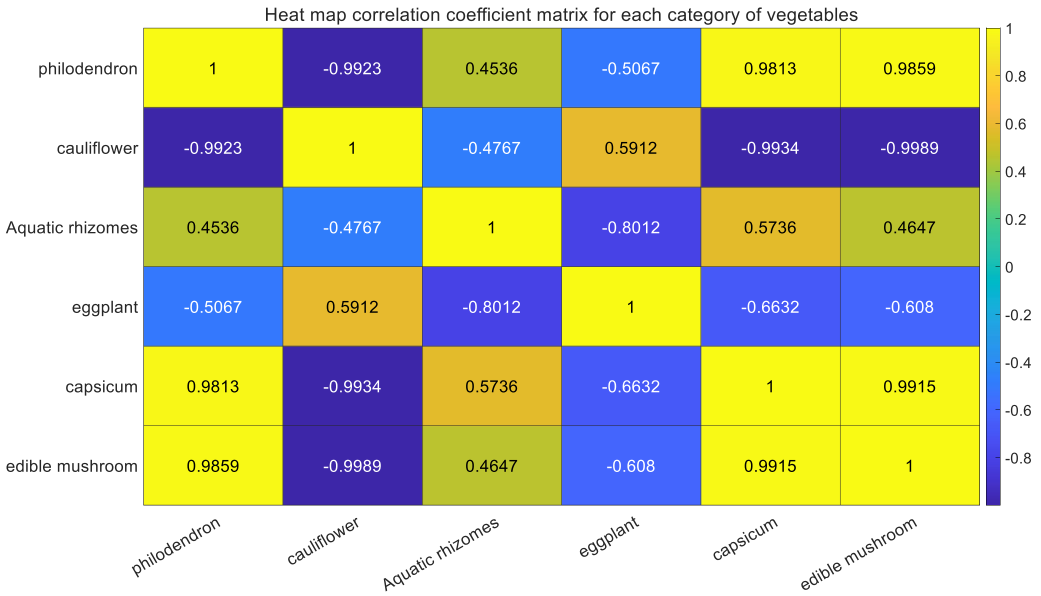

Figure 3. Heat map correlation coefficient matrix for each category of vegetables

Figure 3 demonstrates the correlation between the categories of vegetables, and we used the Pearson correlation coefficient. Pearson correlation coefficient, this design from Karl Pearson, initially defined by statisticians, is mainly used to measure the degree of linear correlation between the variables, covering the range of -1 to 1, designed to reflect the trend and strength of change between the variables, a positive value indicates a positive correlation, while a negative value corresponds to a negative correlation, and if it is 0, it means that there is no correlation between the variables. Larger values represent stronger correlations [7]. Through Figures 4 and 5, it can be known that chili peppers and edible mushrooms are the strongest positive correlation, which may be due to the fact that local people are accustomed to using chili peppers and edible mushrooms together when cooking, which is in line with local tastes.

3. Vegetable commodity replenishment and pricing strategy

3.1. Establishment and Solution of Multiple Linear Regression Model

Since the relationship between the total sales volume of each category and the sales price and wholesale price is analyzed, six multiple linear regression equations need to be solved. Here we take the flower and leaf category as an example, take the total sales volume of each month as the dependent variable, the total sales price of each month and the total wholesale price of each month as the independent variables, and establish the following multiple linear regression equations:

$$Y_1 = a_0 + a_1 x_1 + a_2 x_2 + b \tag{1}$$

Where, $a_i(i = 0,1,2)$ is the parameter to be solved for the regression equation and b is the error term obeying a normal distribution $N(0,\sigma^2)$.

After importing the processed data into Matlab, it was solved by the regress() function for solving the multiple linear regression equation, and the fitting effect was analyzed based on the values of R^2,F,P and S^2 as well as by plotting the residuals. Based on the first fitting, the solution is obtained:R^2=0.5282,F=18.4984,P=0.000004,S^2=0.0102.

According to the value of R^2, we find that the fitting effect is not very good, the plot of residuals found that the residual values of the 24th, 26th and 27th months did not pass the zero line, we need to eliminate such anomalies to deal with the elimination of these three months of abnormal values after the second fitting. Analyzing the reasons for the outliers, it is possible that the extreme weather in that month or other seasonal factors affecting the sales of vegetables lead to a large difference between the sales in these months and the other months.

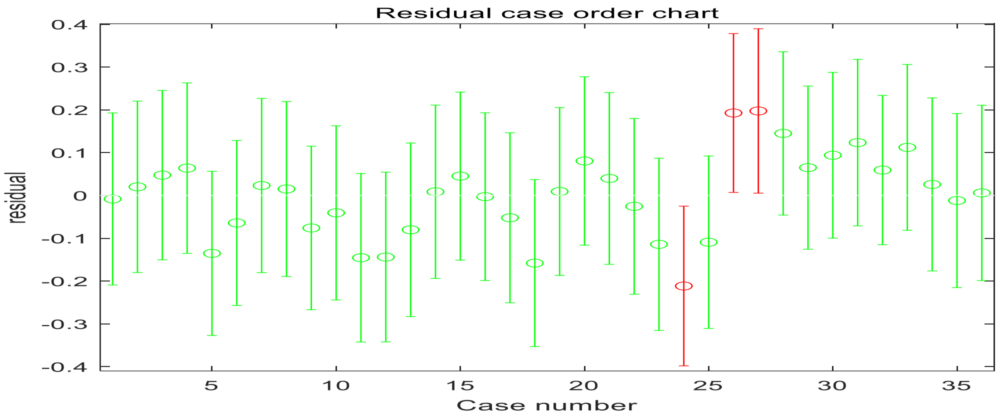

Figure 4. Plot of the residuals of the first fit

After five fits there are no outliers, solving for $R^2 = 0.8047$, $F = 22.3714$, $P = 0.000002$, $S^2 = 0.0050$,at this time the fit R^2 is larger and the value of F is large,plotting the residuals found that at this time all the remaining months of the residual values are through the zero line, no outliers appear, the fit ends here.

According to the last fitting results, we get all the values of the parameter a_i to be solved obtained by Matlab solving as follows:a0=0.7334, a1=-0.0929, a2=0.0942.

In summary, we can finally get the expression of the multiple linear regression equation between the total sales volume Y_1 of the flower and foliage category in each month and the total sales price X_1 in each month and the total wholesale price X_2 in each month as follows:

$$Y_1 = 0.7334 - 0.0929 X_1 + 0.0942 X_2 \tag{2}$$

where, $R^2 = 0.8047$, $F = 22.3714$, $P = 0.000002$, $S^2 = 0.0050$,the multiple linear regression equation is well fitted by the value of R^2.

Based on the R^2 and other test values of all the above multiple linear regression equations, it can be seen that all the equations are fitted better and can better quantitatively respond to the total sales volume and sales price of each category.

In order to show the correlation between all the above multiple linear regression equations and their corresponding variables in a more intuitive and three-dimensional way, we plotted the three-dimensional images of all the above multiple linear regression equations in the three-dimensional coordinate system (as shown in Figure 5). It is easy to observe and get the conclusion, for example, the total sales volume Y_1 is negatively correlated with the total sales price X_1 and positively correlated with the total wholesale price X_2.

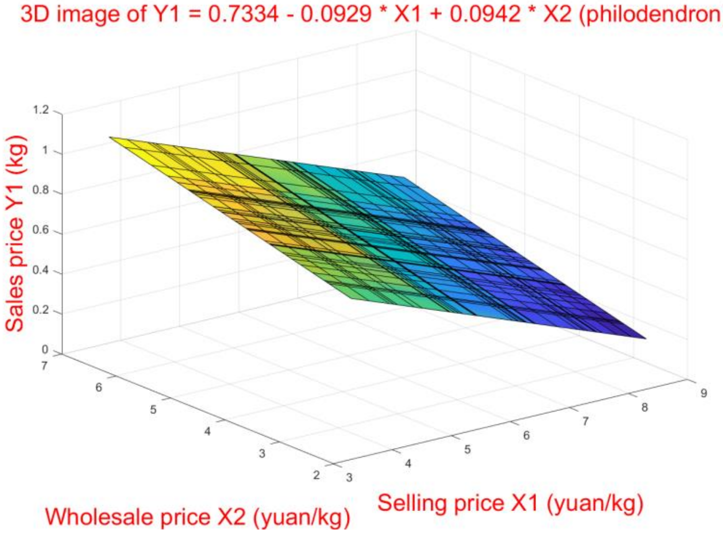

Figure 5. 3D image of Formula (2)

3.2. Total daily replenishment and pricing strategy for the coming week

Pricing strategy needs to determine the total daily replenishment before proceeding, there is a sequential relationship in. Since in the above, we have solved the multiple linear regression equations for all categories and selling prices, wholesale prices, so here we consider the use of the above multiple linear regression equations to carry out a regression prediction model analysis, the use of existing data to predict the total daily sales in the coming week, using the known attrition rate data, the results of the formula $\text{Predicted total sales volume} /(1-\text{Loss rate})$ as the daily replenishment [8], and then Using the "cost-plus pricing" method, a reasonable pricing strategy is given to maximize the revenue.

3.2.1 Modelling and solving

We will filter the data into Matlab, in order to simplify the problem, we will be the daily total sales price of each category is set to $S_{ij}(i=1,2,...,7;\ j=1,2,...6)$; at the same time, we also need to count the number of records sold each day in each category $C_{ij}(i=1,2,...,7;\ j=1,2,...6)$. Here we will be 3 years of the same date of the total number together, so we need to find the final average sales price needs to be divided by 3, the following other calculations are also the same here. Use the following formula to get the final average sales price per day for each category Avr（ij）:

$$Avr_{ij}=S_{ij}/C_{ij}/3;(i=1,2,...,7;\ j=1,2,...6) \tag{3}$$

Following the same methodology, for the wholesale price data in Annex 3, we were able to use Matlab to calculate the total wholesale price for each category $B_{ij}(i=1,2,...,7;\ j=1,2,...6)$ and the number of wholesale records per day for each category D_{ij}. The following formula is used to obtain the final average wholesale price per day for each category Buy_{ij}:

$$Buy_{ij} = B_{ij} / D_{ij} / 3; (i = 1,2,...7; j = 1,2,...6) \tag{4}$$

After obtaining the average sales price and average sales price of each category per day, we will predict the total sales volume of each category per day, using the multiple linear regression equations established above about the total sales volume of each category regarding the sales price and wholesale price, we will substitute the average sales price and average wholesale price of each category per day into the corresponding regression equations. At the same time, in order to simplify the model analysis and solution, we consider that the number of sales records of each category per day is equal to the average number of sales records per day, so we can predict the total sales volume of each category per day by the following formula $Sale_{ij}$:

$$Sale_{ij} = Y_j C_{ij} / 3 (i = 1,2,...,7; j = 1,2,...6) \tag{5}$$

Where, $Y_j = a_0 + a_1 Avr_{ij} + a_2 Buy_{ij}$, represents the average daily sales volume per sales record for category j.

In the above steps, we can get the total daily sales volume of each category, then we import the Matlab data and calculate the average wastage rate of each category according to the method in the idea analysis $C_j, j = 1,2,...6$, and use the following formula to get the final total daily replenishment of each category Add_{ij}:

$$Add_{ij} = Sale_{ij} / (1 - c_j); (i = 1,2,...7; j = 1,2,...6) \tag{6}$$

After importing the data into Matlab, the calculation is performed step-by-step according to the above formula, and the programming can be solved to obtain the predicted total daily sales of each category and the average wastage rate of each category. Finally, using the formula (4), the total daily replenishment of each category can be predicted, as shown in Table 1 below, the total daily replenishment matrix rows represent 7 days a week, columns represent 5 different categories, for example, row 2, column 1 represents the total replenishment of the second day of the flower and leaf category.

Table 1. Forecasted daily replenishment for each category for one week

Kind / dates	Philodendron	Cauliflower	Aquatic rhizomes	Eggplant	Chilli	Edible mushroom
1 July	268.7238	45.6235	23.5953	28.9895	121.7711	104.9295
2 July	281.9073	51.3205	32.6476	32.5722	135.4398	117.4000
3 July	307.5201	57.6388	56.1055	40.2539	150.1112	148.0608
4 July	300.2646	55.7485	33.3946	41.6474	146.3063	130.1995
5 July	301.4046	62.7564	33.0941	46.4314	159.5620	144.0983
6 July	283.3956	49.5317	28.0990	36.0120	129.7216	108.2826
7 July	248.6098	55.0893	30.4925	31.3036	112.9160	108.2826

3.2.2 Pricing strategy development

For the pricing strategy that can maximize the revenue of supermarkets, after reviewing the relevant literature, it is known that the pricing strategy of "cost-plus pricing" can maximize the revenue of supermarkets [9-10], and the formula for cost-plus pricing is as follows:

$$P_{ij} = (1 + r_{ij}) * X_{ij} \tag{7}$$

Where P_{ij} is the pricing, r_{ij} is the markup rate and there are $r_{ij} = (\text{Selling price - Purchase price}) / \text{Purchase price}$ and X_{ij} are the costs. In order to simplify the solution of the model, we make reasonable assumptions. In this problem, we assume that the considered cost X_{ij} is related to both the average wholesale price Buy_{ij} and the average wastage rate C_j and satisfies the following relationship:

$$X_{ij} = (1 + C_{ij}) * Buy_{ij} \tag{8}$$

With all the data known earlier, we can derive the cost per day for each category X_{ij}, for the markup rate r_{ij}, we consider calculating the different markup rates per day for each category with the following formula:

$$r_{ij} = (Avr_{ij} - Buy_{ij}) / Buy_{ij} \tag{9}$$

Based on Eq. (6) and Eq. (7), and the data that has been imported in Matlab, we calculate the cost per day for each category, and then calculate the cost-plus pricing matrix as shown in Table 2 below:

Table 2. Daily cost-plus pricing for each category for one week

Kind / dates	Philodendron	Cauliflower	Aquatic rhizomes	Eggplant	Chilli	Edible mushroom
1 July	0.5943	0.8027	0.2835	0.3998	0.4453	0.8266
2 July	0.6355	0.8137	0.5031	0.6339	0.7238	1.2875
3 July	0.6189	0.8079	0.5462	0.5212	0.5737	1.2987
4 July	0.6569	0.8779	0.3788	0.5928	0.6053	1.2069
5 July	0.5487	0.7791	0.7269	0.7926	0.4383	1.0621
6 July	0.6188	0.7515	0.4638	0.7943	0.8230	1.4101
7 July	0.5742	0.6826	0.5124	0.7549	0.8388	1.0091

Ultimately, based on the contents of Tables 6 and 7, the final pricing strategy can be obtained by substituting Equation (5) as shown in Table 8 below, where each element of the matrix P_{ij} represents the pricing ($/kg) for day i of the j category.

Table 3. Daily pricing for each category for one week

Kind / dates	Philodendron	Cauliflower	Aquatic rhizomes	Eggplant	Chilli	Edible mushroom
1 July	2.5179	3.8416	6.0260	1.9671	3.6430	4.4976
2 July	2.4220	3.6810	6.3501	1.8656	4.1310	5.0927
3 July	2.4645	3.6471	4.8941	2.0742	3.7400	5.3125
4 July	2.5925	3.8451	5.4902	2.0340	3.8908	5.2441
5 July	2.4206	3.5899	5.4163	2.1184	3.8279	4.8314
6 July	2.5216	3.6049	4.9987	1.9951	4.2061	5.4717
7 July	2.4845	3.3822	5.1580	1.8219	4.6207	5.2687

4. Conclusion

In this paper, multiple linear regression models were used to predict the sales volume of each category or individual item of vegetables, and the regression results were fitted multiple times using residual plots. For the relationship between total sales by category and cost-plus pricing, multiple linear regression equations were established to more accurately predict the sales volume of each category or individual item of vegetables on a given day. For the relationship between the total sales of each category and cost-plus pricing, multiple linear regression equations are established, such as for the total daily replenishment, using the existing data to obtain the average sales price and average wholesale price, Substitute the multiple linear regression equation established above to obtain the predicted total sales volume, and through the daily replenishment of total predicted total sales volume / (1 loss rate) to obtain the daily replenishment of total daily replenishment of 7 days. The data obtained in this paper that help superstore replenishment and pricing decisions are inventory data, transportation cost data, product quality data, customer taste preference data, and selling time data.

Regarding the improvement, the analysis of sales volume distribution law and interrelationships can be further improved. For example, the analysis of sales volume distribution pattern can be analyzed from the following dimensions: time distribution, product distribution and price distribution, the model can be considered to use a nonlinear programming model, with daily replenishment and daily pricing of goods as the decision variables, and the total profit of a week as the dependent variable, and then consider the use of dynamic programming or the gradient descent method to solve, and then finally, the results of the sensitivity analysis and robustness analysis.

Reference

[1] Yan Haowei,Mu Yueying. Vegetable industry monitoring and early warning system construction research based on vegetable market price monitoring and early warning[J]. Vegetable,2023(05):1-10.

[2] Liu Jinyan. Analysis and Countermeasure Suggestions for Achieving Tight Balance of Supply and Demand in Vegetable Market in the Context of Large-scale Emergencies[J]. Business Economy,2023(08):101-103+196.

[3] Ou Yanfeng. Research on joint replenishment and cost sharing strategy of perishables with time-varying deterioration rate[D]. Southeast University,2022.

[4] Shao Juping, Cao Qian, Shen Minyan et al. VRP model and algorithm with time window for fresh produce distribution[J]. Industrial Engineering and Management,2015,20(01):122-127+134.

[5] Zhai Zhihong,Jiang Minxing,Chang Chunying. Shock effect of precipitation on vegetable prices--Taking Guangzhou as an example[J]. Resource Science,2021,43(02):304-315.

[6] Liu Tianyu. Research on the optimization problem of warehousing and distribution mode of third-party platform for community group purchase fresh food under time window constraint[D]. Zhengzhou University, 2019.

[7] Cheng Juanjuan. Empirical research on the relationship between research and teaching in colleges and universities--analysis based on Pearson's correlation coefficient[J]. China Colleges and Universities Science and Technology,2022

[8] RAN Wenwen, LIU Huijuan, LIU Yujie, ZHANG Zhe. Research on fresh food supply strategy based on JMI[J]. China Market,2017,(15)

[9] An Kui. Research on cooperative optimization of M fresh supermarket supply chain inventory[D]. Guizhou University,2022

[10] Mao Lisha. Research on pricing strategy and production and marketing model of vegetable wholesale market under the perspective of supply chain[D]. Hunan: Central South Forestry University of Science and Technology, 2022.