Frontiers in Economics and Management Volume 6 Issue 11, 2025
ISSN: 2692-7608 DOI: 10.6981/FEM.202511_6(11).0019

Vegetable Purchase and Pricing Strategy Combined with ARIMA, K-means, LSTM and Optimal Linear Programming

Ziwei Li

Department of Statistics and Data Science, Faculty of Science and Technology, Beijing Normal-Hong Kong Baptist University, Zhuhai, 519087, China

Abstract

The shelf life of vegetables is short, and with the serious loss of time, it is important for supermarkets to make reasonable replenishment decisions and pricing. Based on the quantity control of single items and the minimum display quantity of each single item, this paper constructs an optimal linear programming model with the ultimate goal of maximizing the benefits of supermarkets and meeting the market demand. Based on ARIMA model, K-means clustering analysis, LSTM model training validation, brought into linear regression analysis to establish a prediction model. It is expected that the compound model will be widely used in supermarkets to make purchase decisions and pricing decisions.

Keywords

Linear Regression Analysis; Optimal Linear Programming; ARIMA Model; K-means Clustering Analysis; LSTM Model.

1. Introduction

Because the fresh-keeping period of vegetable products is generally short, supermarkets must make replenishment decisions when the specific products and purchase prices are unknown, and a reasonable sales mix is very important for supermarkets. Vegetable pricing generally adopts the method of "cost plus pricing" (Shang et al., 2025)[1] and supermarkets will sell goods with poor transportation loss and poor quality at a discount. In order to maximize the benefits of supermarkets, supermarkets should make reasonable pricing decisions and replenishment decisions. Therefore, a mathematical model is established based on historical data and actual situation.

2. Data Source, Symbol and Formula Description

The data set was obtained from the database of China Society of Industrial and Applied Mathematics in 2023 (China Society of Industrial and Applied Mathematics, 2023), with a sample size of about 870,000[2].

Table 1. Symbolic Description

symbol	Description	Unit
x	Replenishment quantity of vegetable category I (kg)	Kilogram
p	Wholesale price of the first vegetable category (yuan/kg)	Yuan/kg
c	Cost component of vegetable category I (percentage)	Percentage
d	Forecast quantity of demand for the first vegetable category (kg)	Kilogram
l	Wastage rate of vegetable category I (percentage)	Percentage
s	Stock of vegetable category I (kg)	Kilogram

ISSN: 2692-7608 DOI: 10.6981/FEM.202511_6(11).0019

The cost plus price uses the following formula:

$$\text{Product target price} = \text{cost base} + \text{percentage} \tag{1}$$

$$\text{Price} = \text{unit cost} * (1 + \text{Addition rate}) \tag{2}$$

$$\text{Unit cost} = \frac{(\text{Total fixed cost} + \text{Total variable cost})}{\text{sales volume}} = \text{unit fixed cost} + \text{unit variable cost} \tag{3}$$

$$\text{Addition rate} = \frac{(\text{selling price} - \text{purchase price})}{\text{purchase price}} * 100\% \tag{4}$$

3. Model Assumptions

(1)The inventory of vegetables is 0 at time t, and in the process of replenishment, vegetables begin to lose once they enter the inventory;

(2)Replenishment occurs instantaneously, the delivery time is ignored and the supermarket is not allowed to be out of stock, and the replenishment time is fixed without delay;

(3)The market demand of vegetables follows the demand price theory model in economics (Nagurney and Besedina, 2023)[3], and the market demand of vegetables is negatively correlated with the sales price of vegetables at a certain moment.

4. Establishment and Analysis of the Model

4.1. Analyze the Possible Relationship between Different Items of Vegetable Products

In order to improve the reliability and validity of the data, the data are cleaned and pretreated, and the reliability and validity test are carried out to get a truly effective data set. There are two main methods to analyze it, which are time series prediction model and K-means cluster analysis (Hui and Zhao, 2025) [4]. Using time series to sort out the data according to its actual occurrence and predict the future period, using K-means cluster analysis method to carry out K-means cluster analysis and research on all sales volume.

After integrating the data, the total sales volume of various vegetables from July to December 2020 is shown in Table 2.

Table 2. Statistical table of total sales volume of various dishes from July to December in 2020

Date	Sales volume					
	Cauliflower	Mosaic	Peppers	Solanum	Edible fungi	Aquatic rhizomes
July	1522.555	6577.078	2198.049	1365.511	1617.863	332.056
August	1748.658	7262.394	2822.344	1139.393	1636.524	867.18
September	1336.066	5668.902	1992.413	671.692	1643.153	798.931
October	1644.963	6395.312	2680.911	771.967	30000.532	1345.118
November	1943.391	6518.569	1837.705	432.511	3281.028	960.945
December	1235.596	6901.731	1427.439	408.995	3475.546	1896.154

Figure 1 shows the sales distribution of each item. It can be seen that the sales volume of some dishes is generally higher than that of other dishes.

ISSN: 2692-7608 DOI: 10.6981/FEM.202511_6(11).0019

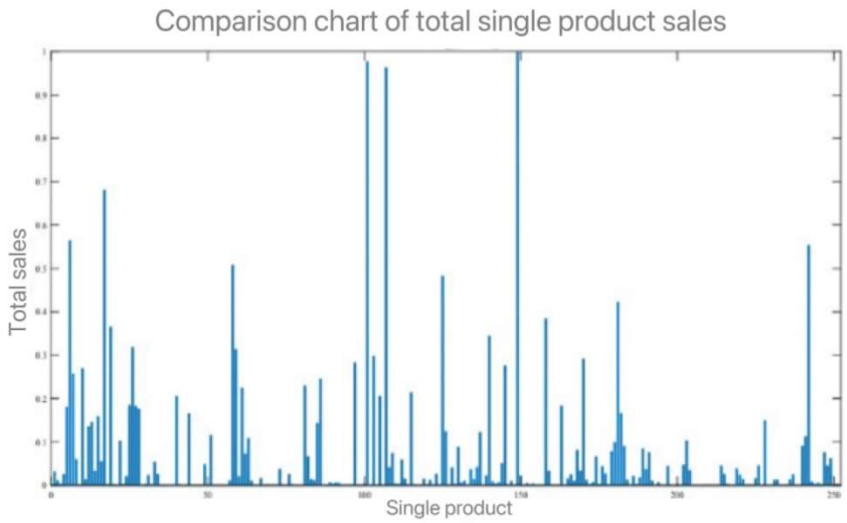


Figure 1. Comparison of sales volume of each single product

Time series simulation, difference and prediction are carried out for the six categories in research. The evaluation results of the model show that there is a certain error between the prediction and the actual situation, but this small error has little influence on the model, which belongs to the acceptable category.

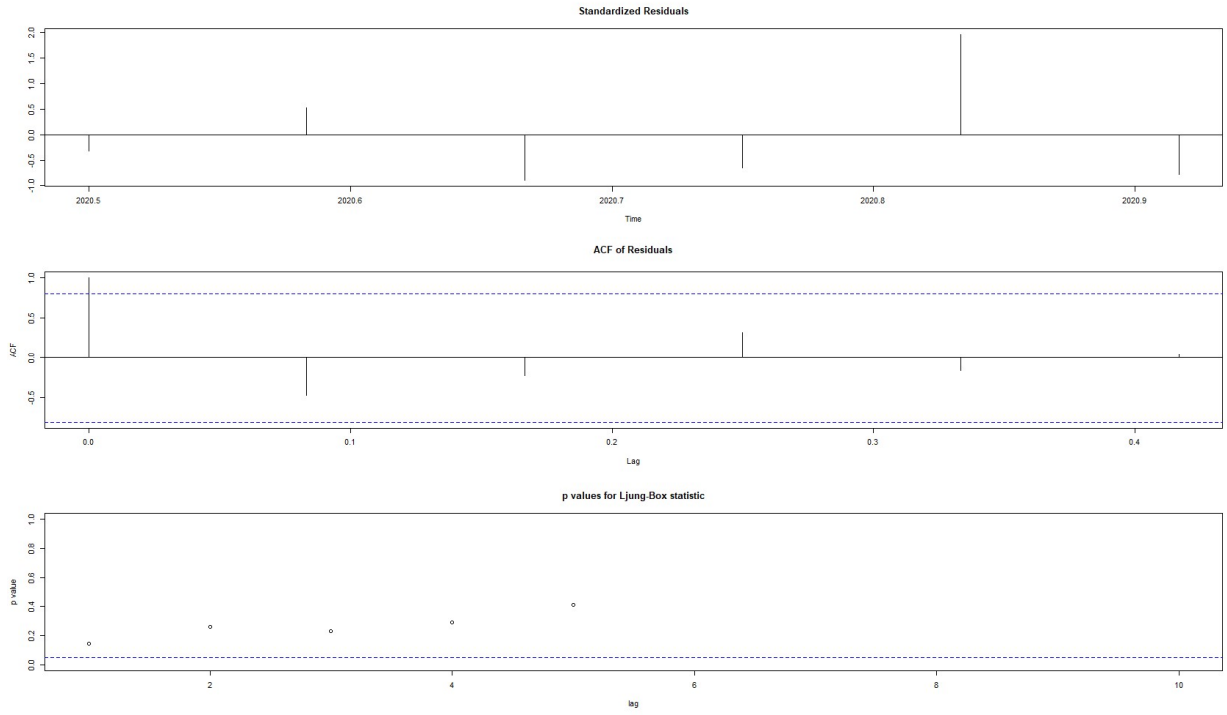


Figure 2. Evaluation result diagram of time series model

In Figure 2, the blue line is the point estimation image, the light blue shaded area near the blue line is 95% confidence interval, and the gray area far away from the blue line is 85% confidence interval. It can be concluded that the model is reliable and stable.

In order to conclude the reliability, K-means clustering was used to analyze the sales volume of various vegetables, and the correlation behind them was studied.

ISSN: 2692-7608 DOI: 10.6981/FEM.202511_6(11).0019

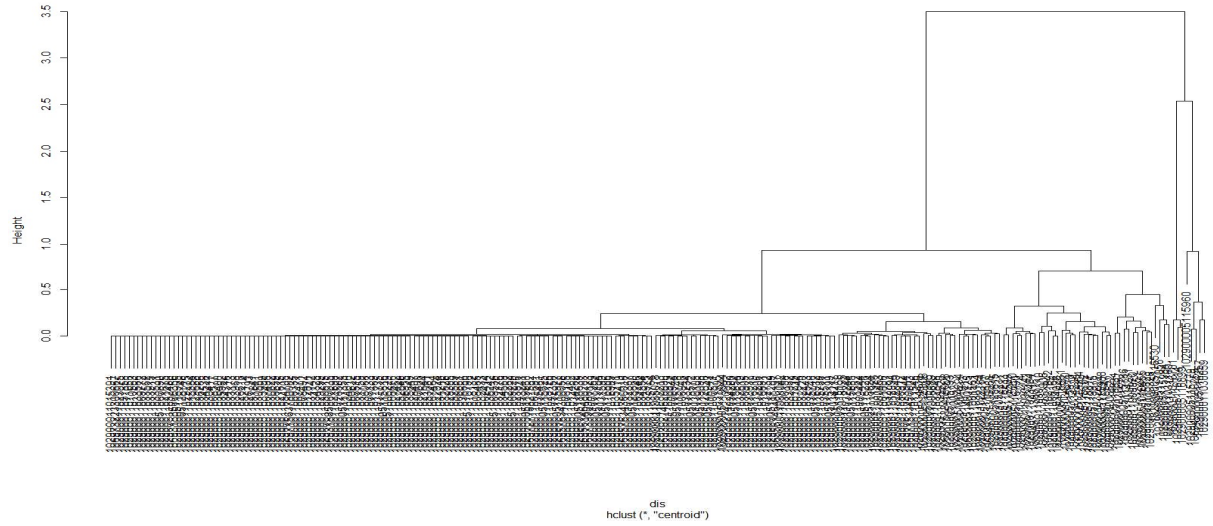


Figure 3. Clustering flow chart

From the flow chart of K-means clustering analysis, we can see that all the sales volume can be clustered into one category only after a few sub-clustering, which proves that there is a strong similarity among tens of thousands of sales volume. Different kinds of vegetables are assigned values, random numbers are generated by using R, two kinds are randomly selected, and two kinds are analyzed by clustering. Mosaic vegetables become one kind alone and have a large difference with others, which proves that mosaic vegetables are in great demand in families and affected by seasonal factors, and become a common kind of vegetables in family vegetables. Compared with mosaic vegetables, the sales gap of other vegetables is not significant, and the frequency of appearing as side vegetables in family vegetables is not as high as mosaic vegetables, so they are attributed to other kinds of vegetables in large class analysis. When buying in supermarkets, the number of mosaic vegetables can be relatively expanded to adapt to the market.

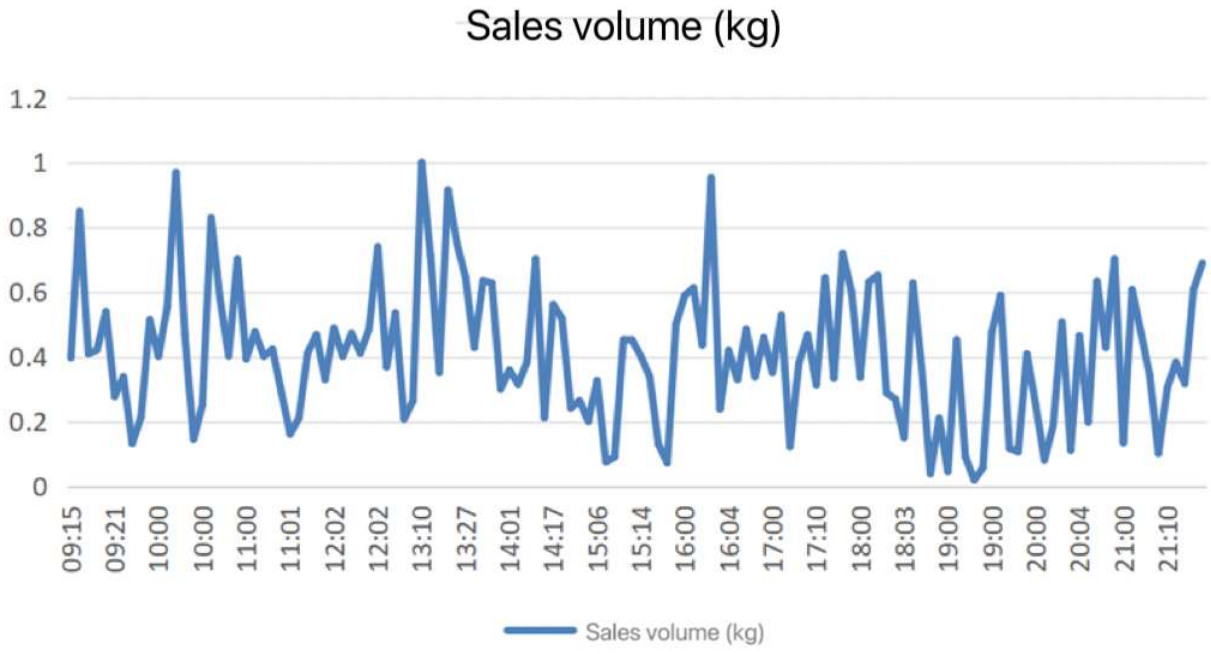


Figure 4. Statistical Chart of Vegetable Sales on July 1, 2020

ISSN: 2692-7608 DOI: 10.6981/FEM.202511_6(11).0019

The following is a box line diagram. From the basic percentile points of the box line diagram image, it can be seen that the bottom image point is 25%, and the top image point is 75%. All the values in this interval are normal values, and those beyond the range are abnormal values. From the diagram, it can be seen that the model has errors, but they are acceptable, which proves the reliability of the model. The results of the third box chart are basically in a certain interval, but there are still some data that are not in the interval. This part of data is presumed to be the data deviation caused by the impact of social environment, which belongs to random disturbance term, and the deviation result is acceptable. As shown in Figure 5.

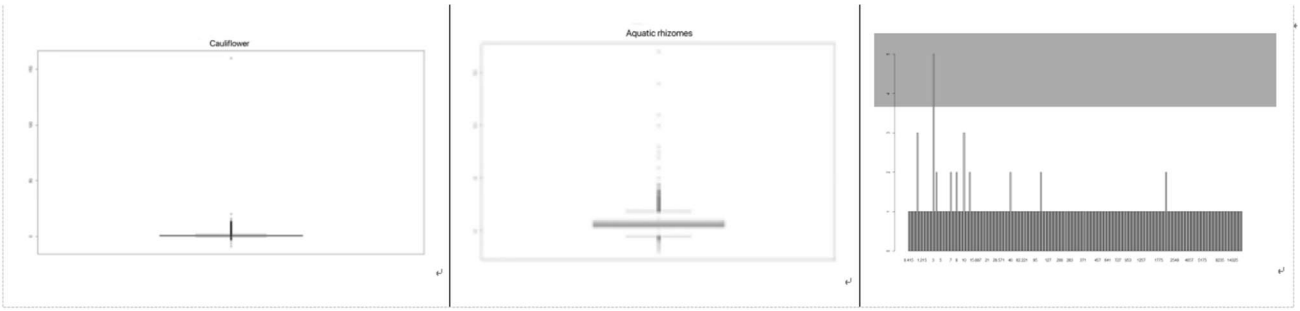


Figure 5. Box line diagram

4.2. Establishment and Solution of Regression Analysis Model between Total Sales Volume and Cost Plus Pricing

LSTM (Long and Short Period Neural Network Prediction) model (Ping. et al., 2024)[5] is a predictive modeling technique to study the specific dependencies between targets and predictors, which is mostly used in time series models. It is divided into linear regression, logistic regression and polynomial regression. After regression analysis of the data, the linear regression equation between total sales volume and cost plus pricing is obtained, and then the replenishment quantity and suggestions in the next week are predicted by using the model. Then import the data to predict the replenishment quantity, and the pricing strategy is calculated according to the formula.

Combined with the function condition of regression equation, it is brought into LSTM (Long and Short Period Neural Network Prediction). The linear regression equation between total sales volume and cost plus pricing is obtained, which is convenient for subsequent modeling and optimization. Then, according to the regression equation, the total amount of replenishment in the next week is predicted. The more sales, the more replenishment. It is required to make the supply equal to the demand for the maximum profit of the supermarket, and the loss of vegetables is not considered for the time being.

Use the vlookup function (Enget. et al., 2025)[6] Integrate the data to obtain a summary table of the total sales volume and cost plus pricing of each single product.

Drawing the scatter plot between sales volume and cost plus pricing, we can see that there is a certain fitting aggregation relationship, as shown in Figure 7.

Where X represents the total sales volume and Y represents the average pricing, because when the total sales volume is found to be zero, there will always be an average pricing, which conforms to the market sales law. To prevent it from affecting the regression function, X=0 is excluded in the regression, and the regression obtains the function f (x) =-0. 0002013*x+12. 89, where R-square: 0.9822 (the square value of the correlation coefficient between the measured data and the reasoning data approaches 1) RMSE: 3.01 (the smaller the mean square error, the better). It is found that this function is a minus function, and its minus degree is small. It can be concluded that the total sales volume of various vegetable categories is negatively correlated

ISSN: 2692-7608 DOI: 10.6981/FEM.202511_6(11).0019

with cost-plus pricing, which is not obvious, although it has a linear negative correlation (Tong, et al). 2025)[7], but has little impact.

It can be intuitively seen that the sales volume of mosaic vegetables is the largest, as shown in Figure 7.

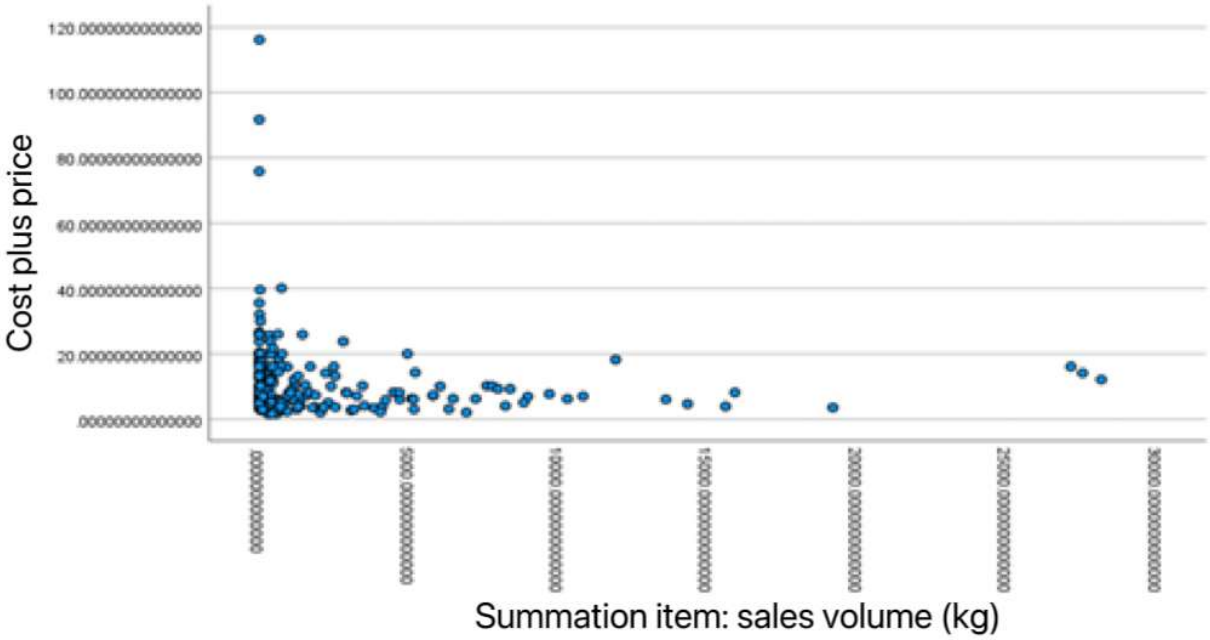


Figure 6. Scatter plot of the relationship between sales volume and cost plus pricing

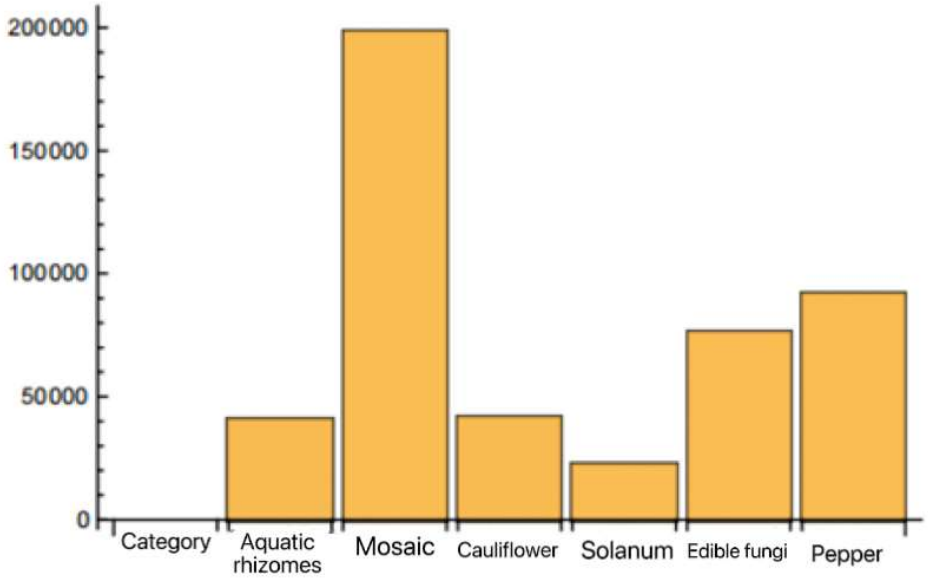


Figure 7. Summary chart of total sales volume of various dishes

The relationship between the major categories is roughly obtained from Figure 8. In order to refine the intuitive diagram, the major categories of vegetables are divided into small single products in detail, and the relationship between them and the total amount is analyzed, as shown in Figure 8.

ISSN: 2692-7608 DOI: 10.6981/FEM.202511_6(11).0019

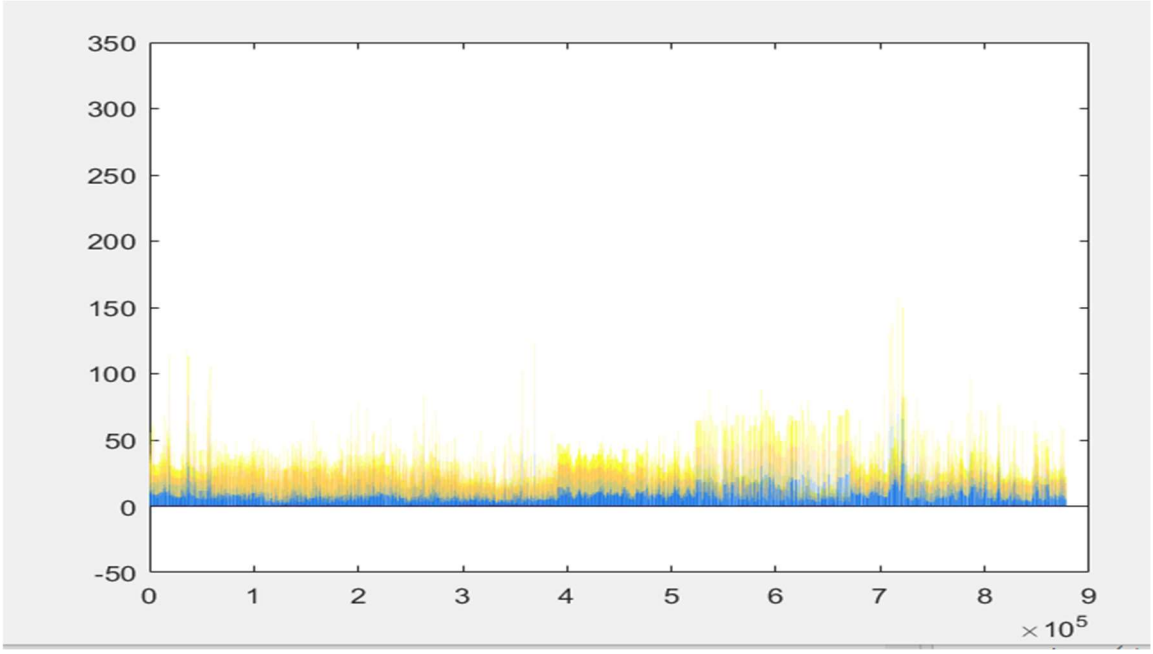


Figure 8. Vertical Stacked Bar Chart

A subaxis matrix (Matsui, et al. (2021)[8] Figure 9, which contains the scatter diagram of X rows and Y columns. Combined with the clustering situation, the total replenishment amount in the next week is predicted according to the regression equation. The more sales, the more replenishment. If you want to maximize the supermarket revenue, make the supply equal to the demand, and bring it into the equation to solve the replenishment amount in the next 7 days.

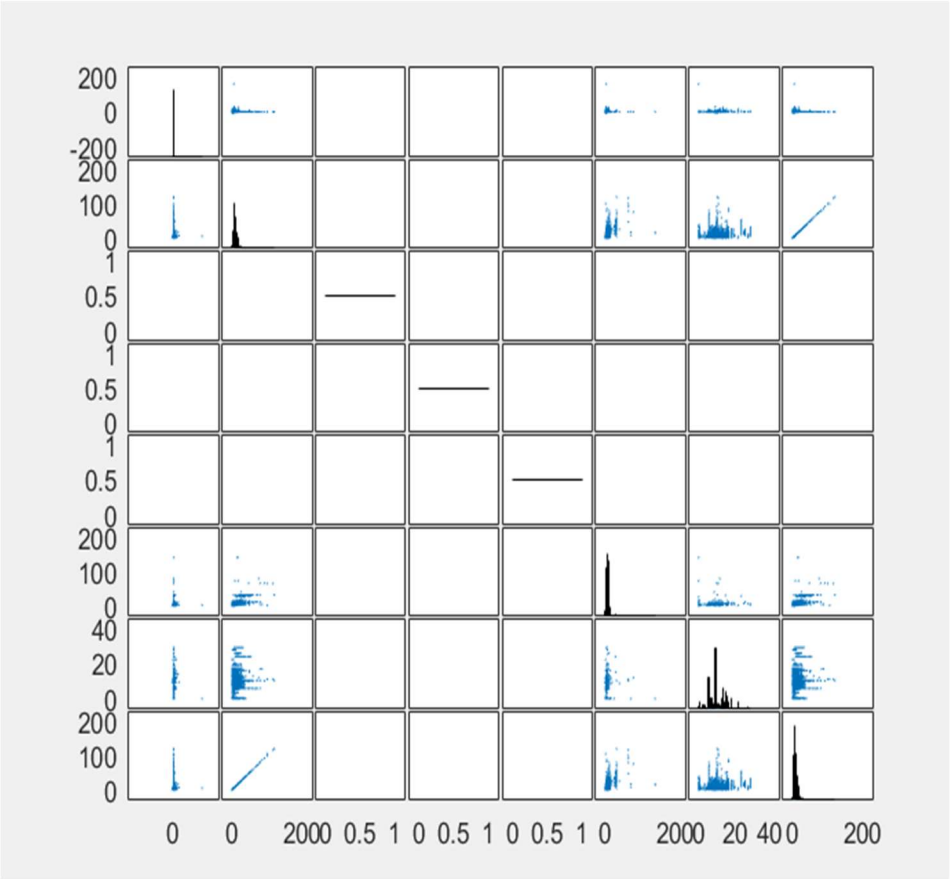


Figure 9. Sub-axis matrix diagram

ISSN: 2692-7608 DOI: 10.6981/FEM.202511_6(11).0019

```
    0.2130    5.2000      NaN      NaN      NaN
    1.0000    4.8000      NaN      NaN      NaN
    1.0000    5.9000      NaN      NaN      NaN
    0.2100   20.0000      NaN      NaN      NaN
    0.2600   12.0000      NaN      NaN      NaN
    0.3880    5.2000      NaN      NaN      NaN
    1.0000    5.9000      NaN      NaN      NaN
    1.0000    3.5000      NaN      NaN      NaN
    1.0000    1.7000      NaN      NaN      NaN
    1.0000    5.8000      NaN      NaN      NaN
    0.5230    5.2000      NaN      NaN      NaN
    0.3490    5.2000      NaN      NaN      NaN
    0.4460    5.2000      NaN      NaN      NaN
    0.0970    5.2000      NaN      NaN      NaN
    1.0000    3.5000      NaN      NaN      NaN
    1.0000    5.8000      NaN      NaN      NaN
    1.0000    2.8000      NaN      NaN      NaN
    0.2520    5.2000      NaN      NaN      NaN

>> size(data)

ans =

      878503           8

>> stem(data,'DisplayName','data')
>> bar(data,'DisplayName','data')
>> plot(data,'DisplayName','data')
>>
```

Figure 10. Replenishment Result Chart

As shown in Figure 10, after each step of data processing, the replenishment quantities in the next 7 days are 0.65490699, 0.42452791, 0.066838272, 0, 0, 0 and 0 respectively.

4.3. Improved Model Establishment and Solution based on Linear Relationship

4.3.1. Constraints

On the basis of linear model, combined with actual factors and historical single product replenishment data, it is deeply discussed that the sales space of vegetable products is limited in general supermarkets, so quantitative constraint inequality conditions are added:

(1) The saleable items are controlled at 27-33 kg;

(2) The order quantity of each single item meets the minimum display quantity of 2.5 kg.

Combined with the problems, it is adjusted to meet the market demand on the basis of the available varieties in the previous week, and the forecast results are obtained.

4.3.2. Analysis and Solution of the Regression Model

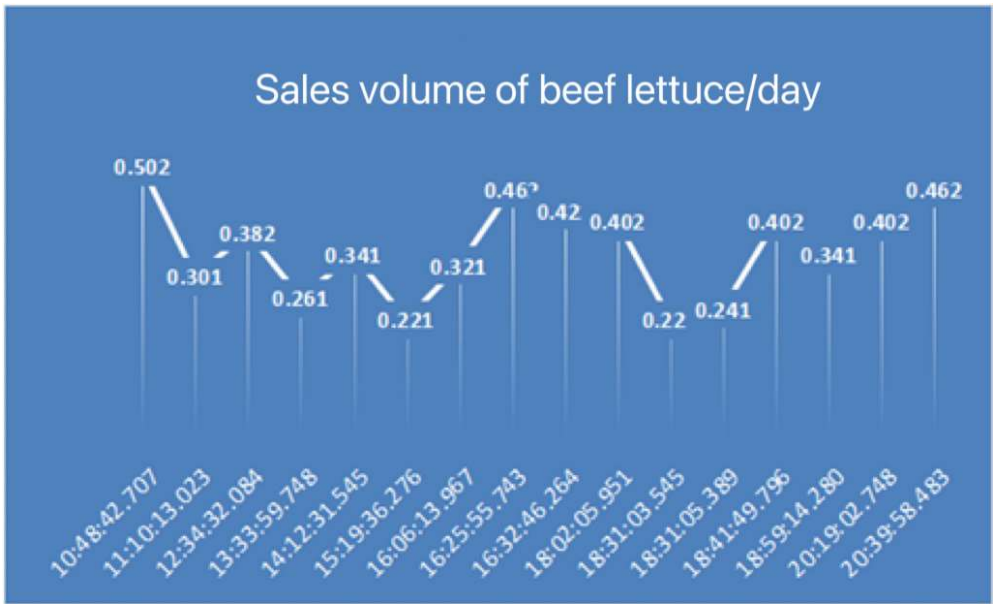


Figure 11. Sales volume of beef lettuce

ISSN: 2692-7608 DOI: 10.6981/FEM.202511_6(11).0019

Figure 11 is the analysis chart obtained by solving the regression analysis model. Taking beef lettuce as an example, the sales volume of beef lettuce in each time period every day is used as the data reference of sales volume.

On the basis of the linear model, the following constraint inequality is added to maximize the total revenue (profit) of the supermarket, which can be defined as sales revenue minus costs, namely:

$$\text{Maximize}\Sigma\ (p_i * x_i - p_i * c_i * x_i) \tag{5}$$

1) The replenishment quantity should not exceed the forecast demand quantity

$$x_i \leq d_i,\ \forall i \tag{6}$$

2) The replenishment quantity should not exceed the inventory quantity to avoid excessive inventory backlog

$$x_i \leq s_i,\ \forall i \tag{7}$$

3) Considering the loss, the sales quantity minus the loss should be equal to the demand forecast

$$x_i - \frac{x_i * l_i}{100} = d_i,\ \forall i \tag{8}$$

$$xi \geq 0,\ \forall i \tag{9}$$

After calculation, the replenishment quantities in the next 7 days are 1.5013416, 1.5854847, 1.4805362, 0.89438140, 1.0913731, 0.43103412 and 0.37696508 respectively.

4.4. Conclusions and Decisions

From table 2, there is a significant correlation between different categories and individual items of vegetable products. Research shows that leafy vegetables have the highest sales proportion, accounting for 42.5%, followed by pepper, accounting for 19.5%.

The K-means clustering algorithm is applied to classify vegetable products, and the research reveals the significant "sales bundling" relationship and differentiated sales distribution law among different kinds of vegetables, which lays a data foundation for the subsequent formulation of targeted sales and marketing plans. At the same time, ARIMA and LSTM models are integrated to build a forecasting modeling system, and based on this, an accurate inventory replenishment plan is formulated, which effectively realizes the dynamic balance between vegetable inventory adequacy and market demand responsiveness, and reduces the risk of inventory backlog and shortage. In addition, seasonal factors have a great influence on sales. In order to explore the relationship between the total sales volume of each vegetable category and cost-plus pricing, ARIMA can verify and predict the results well, and can skillfully manage the balance between sales volume and profitability. In order to maximize profits, LSTM model finds the optimal pricing parameters within the predicted sales range. Then, these results are applied to quantitative inequalities in linear programming to obtain pricing strategies related to profit maximization. The integration and application of ARIMA and LSTM model realizes high-precision prediction of vegetable sales volume and loss rate, provides core support for linear

ISSN: 2692-7608 DOI: 10.6981/FEM.202511_6(11).0019

programming of pricing strategy optimization, and helps supermarkets establish an optimal trade-off mechanism between sales volume regulation and profitability guarantee.

In order to further increase the total revenue of supermarkets, "colorful peppers" is considered to be the most ideal item besides the leafy vegetables with the first sales volume normally replenished according to the replenishment quantity. As the second vegetable in sales volume, it can well influence the sales volume of other kinds of vegetables and meet the principle of consumer psychological relevance in economic theory. It is determined that the replenishment quantity is 33.0 kg in the quantitative inequality, and the price is a quantitative pricing strategy.

Using the analytic hierarchy process to evaluate and analyze the replenishment strategy and pricing decision of supermarkets, it is required to maximize the profits of supermarkets. We should not only get the above conclusions from these massive data, but also consider many factors, such as time, consumers and investors.

(1) Considering the time factor, customers' demand for vegetable categories is quite different, and because of different seasons, the pricing of vegetables will also change. In different seasons, the planting difficulty of each vegetable category is different, so the pricing is different, and the pricing affects the consumption situation of customers. In the corresponding season, dynamic replenishment and pricing according to the seasonal characteristics can better meet the consumption psychology of the public, and it is not easy to overstock vegetables and cause loss of interests.

(2) Considering the consumer factor, different regions of consumers have different demands for vegetable categories. According to the data, Changsha, Nanjing, Lianyungang and other cities have become the mainstay cities of vegetable consumption in recent years, contributing their own strength to promote agricultural development. Shouguang, Shandong Province is the hometown of vegetables in China, and its annual vegetable supply is the largest production place in China. Therefore, the less vegetables lost in Shandong or cities close to Shandong, the lower the pricing. On the contrary, in areas where vegetables are scarce, the amount of vegetables lost is large, the transportation cost is high, and the cost of preservation is different, so the pricing is different. People in different regions have different spending power, so make corresponding replenishment strategies and pricing decisions according to different regions. At the same time, combined with association rules, we can increase sales volume and satisfaction by strategically bundling complementary commodities.

(3) Considering the factor of investors, after market investigation, investors put in products or equipment to keep fresh vegetables, which will greatly reduce the loss rate and solve the loss caused by long-distance transportation of vegetables. In order to solve the problems of vegetable insurance, investors need to solve practical problems, increase investment in scientific research, improve scientific and technological support, make better use of space, improve shopping experience and efficiently manage inventory, so as to maximize benefits.

5. Conclusion

5.1. Model Advantages

In this paper, the data is pretreated when establishing the model, and the data used is effective and the amount of data is huge, so the prediction model obtained is sensitive and the error of the prediction results is small, which can be widely used in many aspects. When the amount of data is small, the error obtained by using this model will be smaller, while when the amount of data is large, it will not deviate too much from reality, and the results are more in line with reality, so the error can be ignored. K-means clustering analysis is used to validate the results, and several methods are used to test the error range. The results show that the error is small and can be ignored.

ISSN: 2692-7608 DOI: 10.6981/FEM.202511_6(11).0019

5.2. Model Disadvantages

The data used in building the model are all the data of a certain supermarket, and the data of many supermarkets can be collected to verify and improve the model in the later stage, so as to make the model more perfect. When building the model, the loss rate is not considered for the time being because of the complexity of the data structure, so it can be improved in the later deepening research.

5.3. Model Improvement

In practice, supermarkets generally take necessary scientific and technological means to protect vegetables in transportation in order to maximize benefits, and also invest scientific and technological means to prolong the life of vegetables, reduce the loss rate, and avoid a large number of vegetables losing benefits due to discounted sales due to poor quality. Therefore, the transportation loss rate and loss rate should be taken into account in the future model improvement.

5.4. Model Generalization

The composite model is sensitive, with small error, large and comprehensive sample size, which can be used in the pricing and purchasing scenario of vegetable supermarkets in real life. It also has a certain reference role for all Internet of Things platforms and e-commerce merchants, and only needs to import their purchasing data and then use the model to predict. Together, these contributions emphasize the combination of modeling analysis and economic theory to cope with the globalization of trade and the changeable, short and fast market.

References

[1] Shang L, Liang S, Jin A T, et al. 2025. Genetic Algorithm-Optimized Regression Tree for Automatic Pricing and Replenishment of Vegitable Products [C]//Proceedings of the 2025 International Conference on Digital Economy and Information Systems (DEIS '25). New York, NY, USA: Association for Computing Machine: 187-193.

[2] China Society of Industrial and Applied Mathematics. 2023 Higher Education Society Cup National Mathematical Modeling Competition for College Students [EB/OL]. (2023-09-07) [2023-09-10]. Http://www.mcm.edu.cn/html_en/node/c74d72127066f510a5723a94b5323a26. Html.

[3] Nagurney A, Besedina E. A Multicommodule Spatial Price Equilibrium Model with Exchange Rates and Non-Tariff Measures for Agri-Food International Trade [J]. Oper Res Forum, 2023, 4 (84): https://doi.org/10. 1007/s43069-023-00262-z.

[4] Hui Z, Zhao K. Research on Network Crime and Security Strategy Based on K-means Cluster Analysis Model [J]. Advances in Computer, Signals and Systems, 2025, 9 (2): DOI: 10.23977/ACSS.2025. 090214.

[5] Ping H, Li Z, Shen X, et al. Optimization of Vegitable Restocking and Pricing Strategies for Innovating Supermarket Operations Utilizing a Combination of ARIMA, LSTM, and FP-Growth Algorithms [J]. Mathematics, 2024, 12 (7): 1054. https://doi.org/10.3390/math12071054

[6] Enget K, Lester M, Riccardi W N. Applied Data Analytics: Experiential Exercises in Revenue and Expense Recognition [J]. Issues in Accounting Education, 2025, 40 (2): 133-147. https://doi.org/10.2308/ISSUES-2023-035

[7] Tong J, Wang T, Wei Q, et al. Serum Albumin is Linear and Negatively Associated With the Risk of All-cause and Cardiovascular Death in Coronary Heart Disease Patients. [J]. Reviews in Cardiovascular Medicine, 2025, 26 (8): 38034. DOI: 10.31083/RCM38034.

[8] Matsui K, Yamamoto K, Tashiro S, et al. A systematic approach to the discovery of genomic findings in clinical practice and research: a proposed framework with colored matrix and decision-making pathways [J]. BMC Med Ethics, 2021, 22: 168. https://doi.org/10.1186/s12910-021-00738-9.