Research On the Sales Law and Replenishment Decision of Vegetable Products Based on Statistical Analysis

Yunting Cai *, Sitan Liu #, Peien Li #

School of Software, Dalian University of Foreign Languages, Dalian, China, 116044

* Corresponding Author Email: 13998457197@163.com

#These authors contributed equally.

Abstract. Through mathematical analysis and correlation model, considering the data of demand analysis and supply analysis, combined with the actual situation, formulate a reasonable replenishment strategy and pricing strategy to help supermarkets increase sales and profit margins and enhance market competitiveness. Aiming at different problems, data collation and cleaning were carried out, and correlation analysis and visualization tools were used to analyze the relationship between vegetable categories and the change law of sales volume [1-3]. For the replenishment problem, a linear regression model and an ARIMA time series model are used to explore the relationship between price and sales volume, as well as future sales forecasts, and corresponding conclusions are drawn [4-6]. Aiming at the problem of replenishment and pricing, a mathematical programming model is established and solved by genetic algorithm, and the optimal replenishment quantity and pricing strategy are obtained [7]. Finally, a comprehensive analysis of the results and the actual situation, the integration of relevant data and reasons for the development of vegetable products replenishment and pricing decisions to provide support [8]. In general, the model and analysis in this paper can effectively improve the economic benefits and competitiveness of the supermarket, and provide a scientific basis for decision-making [9-10].

Keywords: Replenishment and Pricing Strategy, Correlation Analysis, Linear Regression Model, ARIMA Time Series Model.

1. Introduction

1.1. Background

Due to the short shelf life and the deterioration of the product phase over time, most of the varieties of general vegetable commodities in the fresh supermarket cannot be resold the next day if they are not sold on the same day. Therefore, supermarkets usually formulate corresponding replenishment plans from two aspects: historical sales and demand conditions of each commodity. The merchant must make the replenishment decision of each vegetable category on the same day without knowing the specific items and the purchase price. The pricing decision of vegetables generally adopts the ' cost markup pricing ' method, and the goods with poor transportation loss and product phase change are often sold at a discount. Due to the special nature of vegetables, reliable market demand analysis is particularly important for replenishment decisions and pricing decisions.

1.2. Research objective

A thorough analysis of the supply and demand of vegetables can help the supermarket to obtain more benefits at a lower cost while ensuring the quality of vegetables, thus contributing to the operation of the supermarket. From the demand side, there is often a certain correlation between the sales volume of vegetable products and time; from the supply side, the supply of vegetables is more abundant from April to October.At the same time, the limitation of the sales space of the supermarket makes the reasonable sales combination extremely important.

It is necessary to collect data, analyze the distribution and relationship of vegetable sales, and make more reasonable replenishment and pricing decisions, so as to better reduce waste and increase profits.

2. Materials and methods

2.1. Data acquisition and preprocessing

2.1.1 Data acquisition

The following websites have content such as vegetable category data, and data such as vegetable category sales are extracted from the website(http://pfsc.agri.cn/#/priceExponent, https://www.gswinfo.com/, https://www.hanghangcha.com/industry).

2.1.2 Data processing

In this paper, we collect the information of single product code, single product name, classification code, classification name, sales date, scan code sales time, single product code, sales volume (kg), sales unit price (yuan / kg), sales type, discount sales and other information of vegetables.

First of all, the data is sorted and cleaned to eliminate outliers and missing values. It is found that the time of the table data is relatively complete, there are no abnormal outliers, and there is no excessive cleaning. The data is sorted into data sets for preprocessing. According to the daily unit, the design program summarizes the daily sales volume of each category from July 1, 2020, to June 30, 2023.

2.2. Methods to introduce.

2.2.1 Analysis of relationship

Correlation analysis is to analyze the degree of correlation between variables. In this paper, it is analyzed that there may be a certain correlation between different categories of vegetable products, as well as the distribution law and interrelationship of sales volume of various categories and single products of vegetables.

2.2.2 Linear regression model

Linear regression is a statistical analysis method that uses regression analysis in mathematical statistics to determine the quantitative relationship of interdependence between two or more variables. It is widely used. The linear regression model is used to explore the relationship between price and sales (including linear regression fitting, polynomial regression fitting, exponential regression fitting, logarithmic regression fitting, and power function regression fitting).

2.2.3 ARIMA time series model

The full name of the ARIMA model is called the autoregressive moving average model, which is the most common model for time series prediction in statistical models. ARIMA model is widely used in the analysis and modeling of various types of time series data. Using the past observations of the sequence, the future value of the sequence can be extrapolated. This paper uses the ARIMA time series model to predict the sales volume in the next seven days through the training test set and gives the total daily replenishment and pricing strategy of each vegetable category in the next week (July 1-7,2023), and maximizes the benefits of the supermarket.

3. Establishment and solution of the model

3.1. The distribution and relationship of the sales volume of each category and single product of vegetables

3.1.1 Analysis of category sales distribution based on Pearson correlation analysis.

In statistics, the Pearson correlation coefficient is used to measure the correlation (linear correlation) between two variables X and Y, with values ranging from -1 to 1. The calculation formula of Pearson correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \quad (1)$$

The closer the correlation coefficient is to 1 or-1, the greater the absolute value of the correlation coefficient, the stronger the correlation; the closer the correlation coefficient is to 0, the weaker the correlation is. Generally, the correlation strength of variables can be judged by the range of values, as shown in Table.1.

Table.1. Related strength grade

Number range	Degree of relatedness
0.8-1.0	**Strongly related**
0.6-0.8	**strong dependence**
0.4-0.6	**medium degree relations**
0.2-0.4	**low correlation**
0.0-0.2	**Very weak or no correlation**

The Pearson correlation coefficient of each vegetable category was calculated, and the relationship between each category was visually displayed through the heat map. The category-Pearson correlation coefficient heat map is shown below in Figure 1. Among them, mosaics and cauliflowers, peppers and edible fungi showed a strong correlation, while aquatic rhizomes, edible fungi and eggplants showed a weak correlation.Among them, cauliflowers, mosaics, peppers, edible fungi, and aquatic rhizomes can be divided into parts, and these categories have a moderate degree of correlation.the correlation between eggplants and edible fungi and aquatic rhizomes was weak.

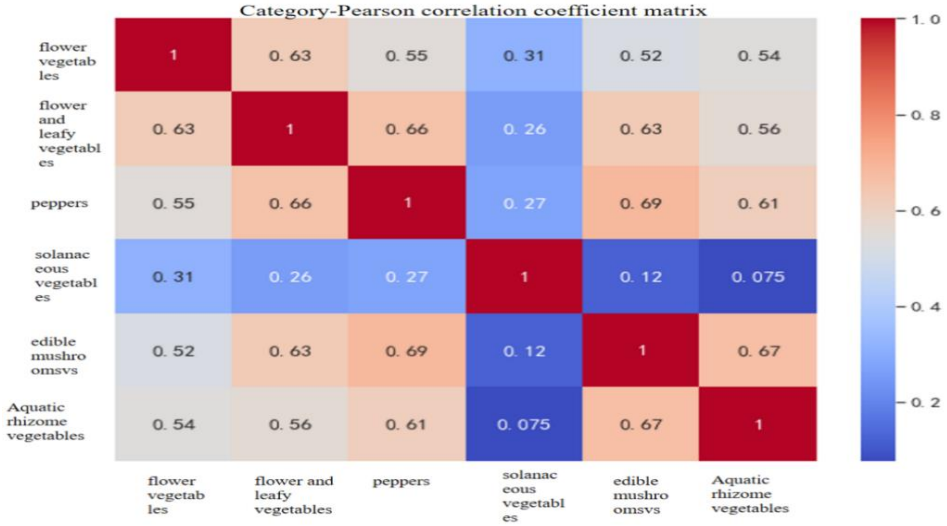

Figure 1. Category-Pearson correlation coefficient heat map

3.1.2 Category distribution based on Spearman correlation analysis

In statistics, Spearman correlation coefficient is a non-parametric index to measure the dependence of two variables. It uses the monotonic equation to evaluate the correlation between two statistical variables. If there is no repeated value in the data, and when the two variables are completely monotonically correlated, the Spearman correlation coefficient is + 1 or − 1.

For samples with a sample size of n, n raw data are converted into hierarchical data, and the correlation coefficient ρ is:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \quad (2)$$

The Spearman correlation coefficient of each category of vegetables was calculated, and the relationship between the categories was visually displayed through the heat map. The category-Spearman correlation coefficient heat map is shown below in Figure 2.

Among them, the correlation between mosaics and cauliflowers, peppers, and edible fungi is relatively strong and positive. The eggplants were positively correlated with cauliflowers, flowers and peppers, but the correlation was weak, while the eggplants were negatively correlated with edible fungi and aquatic rhizomes.

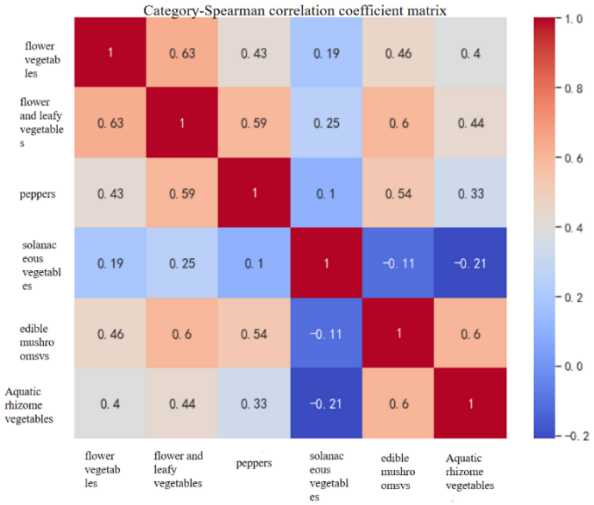

Figure 2. Category-Spearman correlation coefficient heat map

3.1.3 Correlation analysis of category sales based on Kendall 's tau-b correlation analysis.

The Kendall Tau-b coefficient Kendall 's tau-b (Kendall) rank correlation coefficient is used to reflect the correlation index of categorical variables, which is suitable for the case where both categorical variables are orderly classified. Non-parametric correlation tests were performed on the relevant ordinal variables.

The calculation formula of Kendall 's tau-b correlation coefficient is as follows:

$$T_b = \frac{C - D}{\sqrt{T - T_r}\sqrt{T - T_c}} \tag{3}$$

The sales volume of the six categories of vegetables over time is output, and the distribution and interrelation of the categories are explored in a visual way in Figure 3.

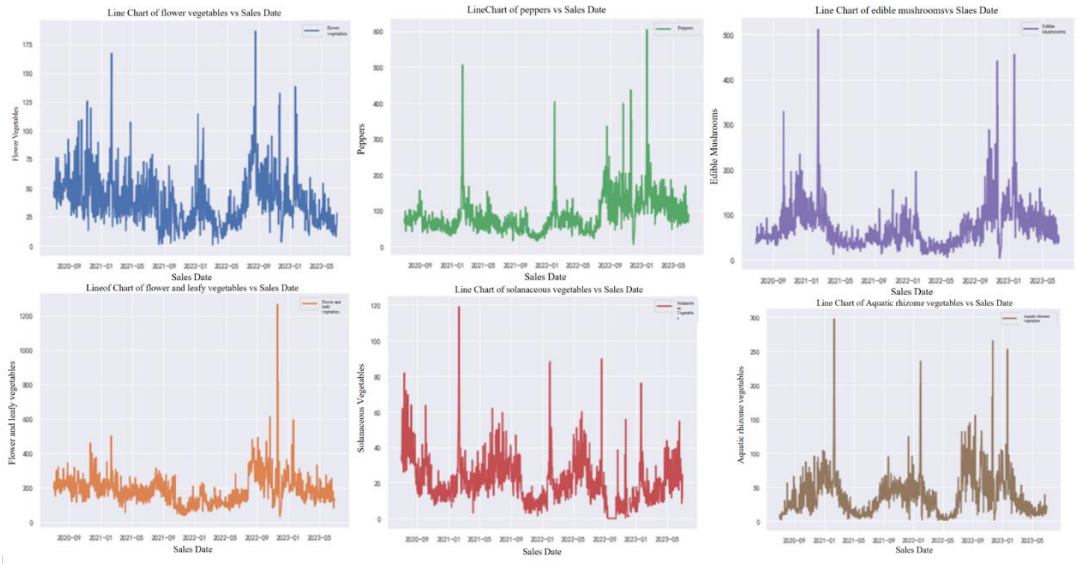

Figure 3. Changes in sales of six categories over time

3.2. Vegetable category replenishment strategy

3.2.1 Solution of regression model and linear fitting

In order to explore the relationship between the total sales volume of vegetable categories and cost-plus pricing, according to the pre-processed data, the total sales volume of each vegetable category was extracted as the dependent variable, and the cost-plus pricing was used as the independent variable. The linear regression model between the corresponding sales volume and cost-plus pricing was established by fitting the model:

$$y = 7.81x + 45.72 \quad (4)$$

In order to evaluate the model more comprehensively, Python linear fitting is used to judge whether the relationship between cost markup price and sales volume is significant, and to confirm whether the model has problems. The specific results are as follows in Figure 4.

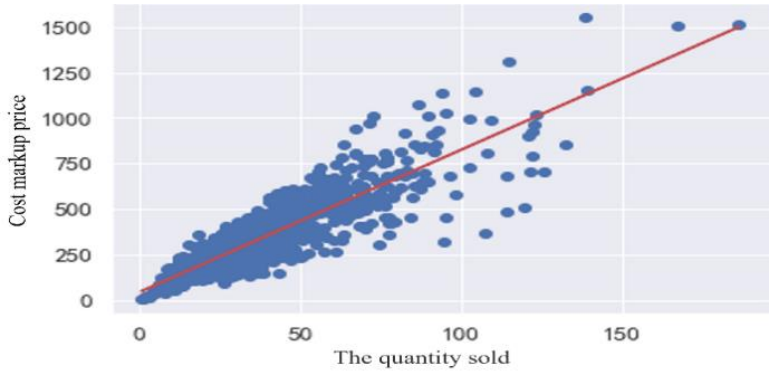

Figure 4. Linear graph of the relationship between cost-plus pricing and sales volume

3.2.2 Establishment and solution of ARIMA time series prediction model

There are three possible cases of replenishment: 1, greater than the expected sales volume 2, equal to the expected sales volume 3, less than the expected sales volume, of which only case 2 is the optimal solution. After analysis, because it is mainly concerned about the future sales forecast, and the sales data has obvious seasonality and trend, a more suitable ARIMA time series model is selected to draw the scatter plot of training value and test value, and predict the sales volume in the next seven days, so as to determine the total daily replenishment and pricing strategy of each vegetable product in the next seven days, so as to maximize the profit of the supermarket.Results are as follows in Table.2 and Figure 5.

Table.2. Category sales volume after data summary

Category/Sales (days)	7.1	7.2	7.3	7.4	7.5	7.6	7.7
flower vegetables	23.11	20.80	19.72	19.23	18.99	18.89	18.84
Leaves and flowers	140.29	143.03	143.79	144.01	144.07	144.08	144.09
Peppers	84.88	86.05	86.58	86.82	86.93	86.98	86.98
Solanaceae	23.33	22.88	22.71	22.65	22.63	22.62	22.61
edible mushrooms	47.95	51.20	52.47	52.96	53.15	53.22	53.25

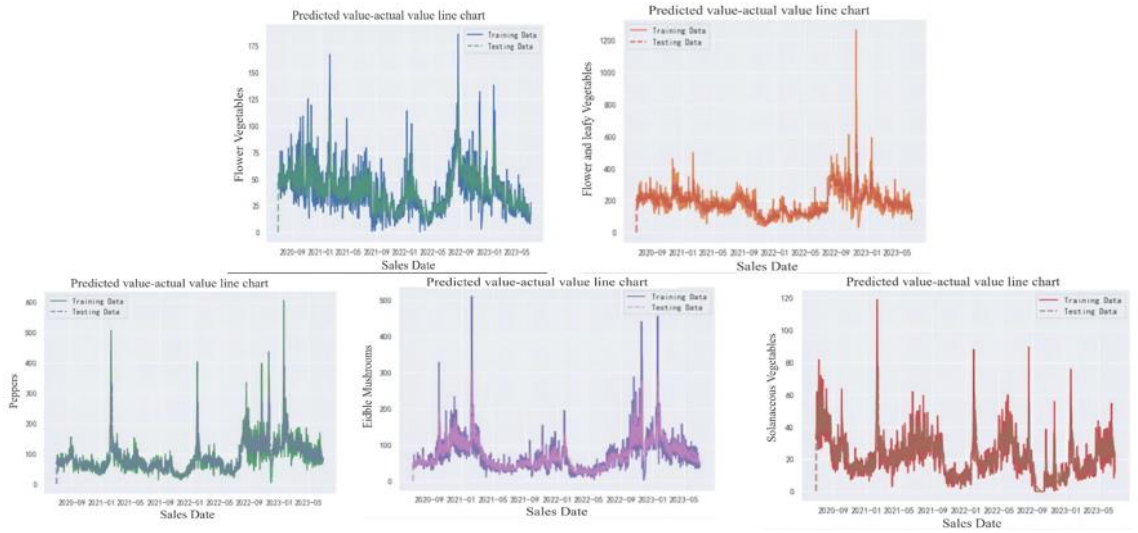

Figure 5. Predicted value-actual value line chart.

4. Conclusions

For fresh supermarkets, vegetable products have the characteristics of short shelf life and deterioration of product phase over time, so they need to be replenished daily to meet the demand. Through the use of mathematical analysis and correlation model, we can comprehensively consider the data of demand analysis and supply analysis, and formulate reasonable replenishment strategy and pricing strategy, so as to improve the sales and profit margin of the supermarket and enhance its competitiveness. Using correlation analysis and visualization tools, we can understand the relationship between vegetable categories and the change law of sales volume, and then make replenishment decisions for different categories. Using linear regression model and ARIMA time series model, we can explore the relationship between price and sales volume and the prediction of future sales and provide decision-making basis for pricing strategy and replenishment quantity. Detailed sales analysis and reasonable replenishment strategy and pricing strategy play an important role in improving the company's over-sales and profit margins and enhancing market competitiveness.

References

[1] Cheng Juanjuan. An Empirical Study on the Relationship between Scientific Research and Teaching in Colleges and Universities - - Based on the analysis of Pearson correlation coefficient [J]. Science and technology in Chinese universities2022(10):46-52.

[2] Wang Xiaoyan, Li Meizhou. Discussion on rank correlation coefficient and Spearman rank correlation coefficient [J]. Journal of Guangdong Industry Polytechnic,2006(04):26-27.

[3] Zhang Weifeng. Statistical characteristics analysis of Spearman correlation coefficient and Gini gamma correlation coefficient [D]. Guangdong University of Technology, 2020.

[4] Liu Yan.Mathematical model of multiple linear regression [J]. Journal of Shenyang Institute of Engineering:Natural Science Edition, 2005, 1(2):128-129.

[5] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and GM Ljung. Time series analysis: forecasting and control. John Wiley & Sons, 2015.

[6] Dong Chunjie. Multi-label classification model based on instance and logistic regression [D]. Nanjing University, 2013.

[7] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T Meyarivan. Multi-objective optimization by genetic algorithms: a review. Proceedings of IEEE International Conference on Evolutionary Computation,1996.

[8] Scientific Platform Serving for Statistics Professional 2021.SPSSPRO. (Version 1.0.11)[Online Application Software]. Retrieved from https://www.spsspro.com.

[9] Lin Fuyong. Logistics cooperation and operation decision-making of fresh e-commerce supply chain under random demand [D]. Jinan University, 2023.05.18.

[10] Mao Lisha. Research on the pricing strategy and production and marketing model of vegetable wholesale market from the perspective of supply chain. Business economic research, (10):139-140, 2019.