Data-Driven Approach to Enhancing Supermarket Vegetable Sales Strategies

Yu Ding [1, *, #], Xi Zhao [2, #], Azez Abdurehim [1, #]

[1] School of Mathematics and Statistics, Kashi University, Kashi, China, 844008

[2] Foundation Department, Southwest Jiaotong University Hope College, Chengdu, China, 610400

* Corresponding Author Email: 17628720172@163.com

[#]These authors contributed equally.

Abstract. China's vegetable commodity production has developed rapidly and made great progress. Due to the short shelf life of most vegetable products, they cannot be sold the next day. Therefore, by studying the distribution law of sales volume and the relationship between categories and pricing, the supermarket can accurately predict the sales situation, reasonably arrange replenishment plans, and formulate pricing strategies. Comprehensive analysis of automatic replenishment and pricing strategies of vegetable commodities is of great significance to the profit of fresh supermarkets. Aiming at the automatic pricing and replenishment strategies of vegetable products, this paper visualizes each item and each category to obtain the distribution and relationship of time, sales volume, and sales frequency of different items and categories. Spearman grade correlation coefficient is used for correlation analysis, and a single linear regression model is established. The relationship between cost profit rate and cost, selling price, and total sales volume of each category is obtained. Then, the ARIMA prediction model is used to predict the automatic pricing and replenishment strategies for each category and each vegetable item during the week of July 1-7, 2023, and July 1, 2023.

Keywords: ARIMA Prediction Model, Python, Spearman Rank Correlation Coefficient, Unary Linear Regression.

1. Introduction

In today's rapidly changing retail environment, supermarkets face increasing competitive pressures and the diversification of consumer demands. Especially in the field of vegetable sales, due to the perishability and seasonal variations of their commodities, supermarkets must accurately predict sales trends and flexibly adjust automatic pricing and replenishment strategies. This requires an efficient, data-driven approach to solving these challenges to maximize profits and increase customer satisfaction.

At present, some simple commodity pricing and replenishment strategies have been unable to meet the needs of the current large market. Chen Long used the unitary linear regression model to take Zhengzhou and other eastern, central, and western cities as samples, which played a certain role in stimulating and promoting urban residents' consumption and economic development [1]. Zhao Jiabao et al. introduced the ARIMA model to forecast the grape production of Turpan City from 2018 to 2020 and obtained a good and stable forecast result for the healthy, stable, and sustainable development of the Turpan grape industry [2]. Wang Hui built ARIMA fiber yield prediction model to analyze and forecast the changing trend of ramie fiber yield and the improvement effect of ramie fiber yield [3]. Li Yuan et al. combined Spearman correlation analysis with time window hierarchy to form a fault detection method that is superior to traditional statistical methods in fault detection [4]. Jia Ke et al. proposed a cascade protection method based on the Spearman rank correlation coefficient and obtained the simulation results of the new energy transmission line and the field recording data to prove the effectiveness of the method [5]. Han Meng analyzed the consumer price index (CPI) and retail price index (RPI) of China from 2005 to 2017 by using a single linear regression model to further understand the relationship between CPI and the retail price index of commodities [6]. Ren Jianying established independent variables and dependent variables through the unitary

linear regression model, used SPSS statistical software to discuss the results of unitary regression analysis, and clearly understood the principle of the parameter estimation method [7]. These studies can be of great significance to the automatic pricing and replenishment strategies of supermarkets in the future.

In this article, we have studied the strategies for maximizing supermarket profits and implementing automatic pricing and replenishment. We have utilized methods such as time statistical analysis, mark-up pricing method, simple linear regression model, and ARIMA forecasting to analyze and statistically analyze the data in the appendix. We have obtained the distribution of vegetable categories, the distribution patterns of sales volume, the relationship between cost profit margin and sales volume, and predicted the maximum profits for each category in the upcoming week. The data source of this article is (www.mcm.edu.cn).

2. Vegetable Sales Visualization Analysis

First of all, the sales data are classified according to the category and single product of vegetables, and the order of the quantity of each vegetable category from large to small is as follows: Flower leaf class (100)> Edible fungi (72)> Capsicum (45)> Aquatic rhizomes (19)> Solanaceae (10)> Cauliflower category (5), that is, the distribution of Flower leaf class, Edible fungi and Capsicum in the vegetable category in the fresh supermarket is more, while the distribution of Aquatic rhizomes, Solanaceae and Cauliflower category is less. Then, the detailed sales data provided were cleaned and organized, and the sales data curve of each category and single product at each time point was obtained, as shown in Figure 1.

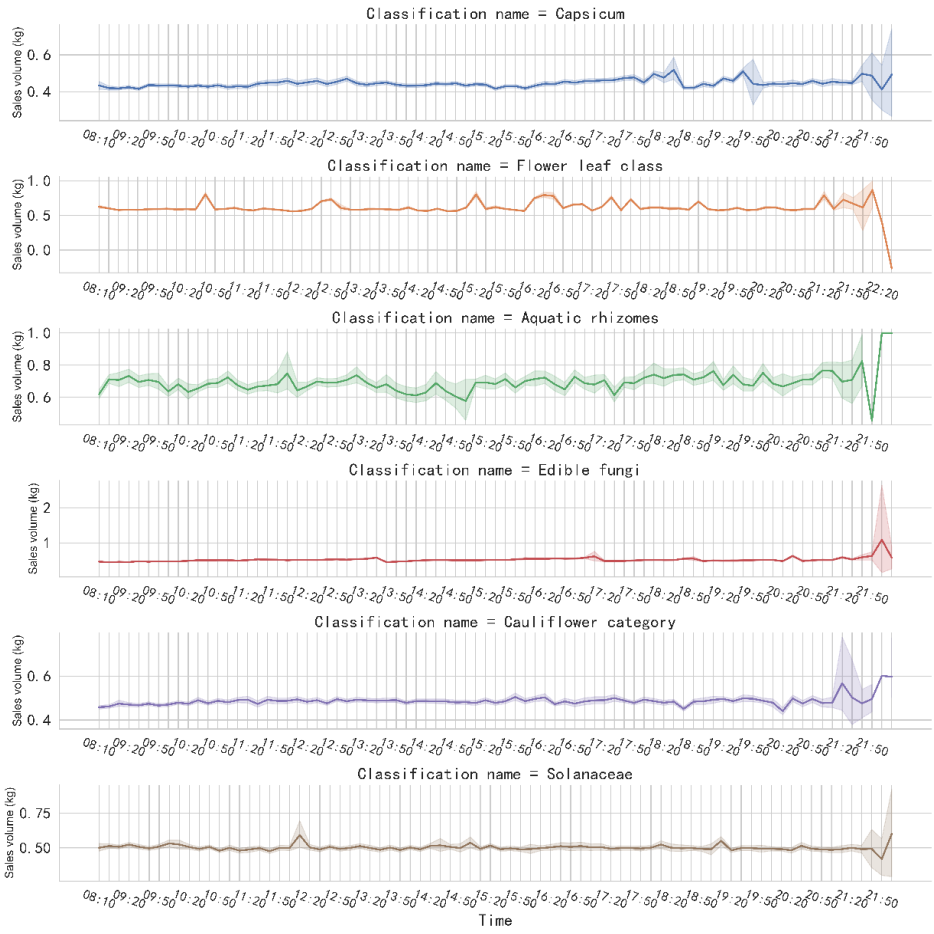

Figure 1. Time distribution pattern of different categories

From the line graph, the sales volume of Capsicum is relatively stable in the period from 08:10 to 18:10 and fluctuates in the period from 18:20 to 22:10, and the overall period is relatively stable. The

sales volume of the Flower leaf class fluctuates from 10:40-11:00, 12:40-13:10, 15:20-15:40, 16:20-18:20, 21:20-22:20, while the sales volume of the Flower leaf class fluctuates during other periods. Aquatic rhizomes fluctuate greatly during the whole period. Edible fungi, Cauliflower category, and Solanaceae were relatively stable in the overall period, but fluctuated in the period from 21:20 to 22:00.

As for the distribution rule of the categories, we selected the top three and bottom three items of the six categories for data cleaning, combined the items from different sources, observed the categories purchased by consumers, and obtained the sales frequency bar chart of different vegetable categories as shown in Figure 2.

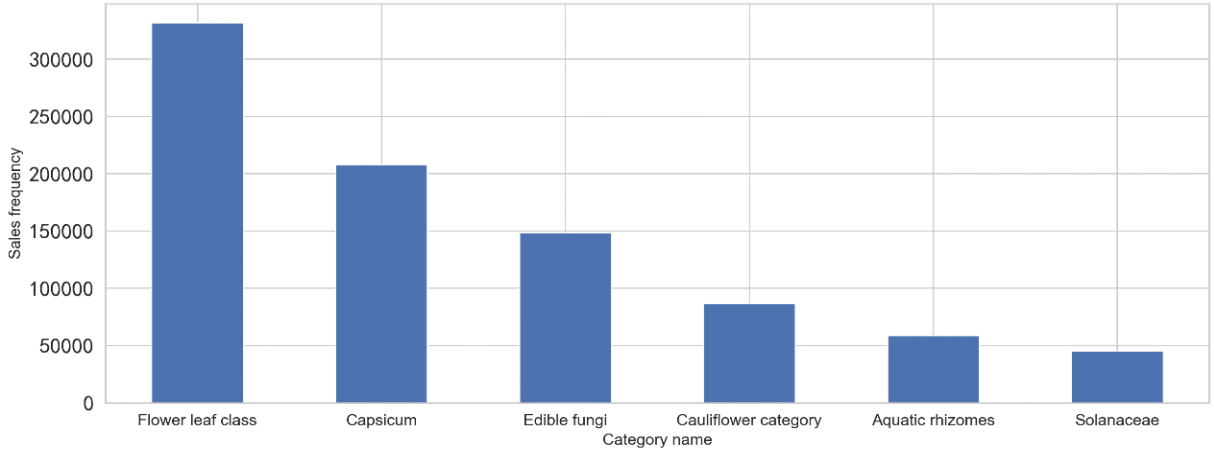

Figure 2. Sales frequency of different categories

Similarly, the distribution patterns of individual products can be analyzed to obtain a bar chart showing the sales frequency of different vegetable categories, as shown in Figure 3.

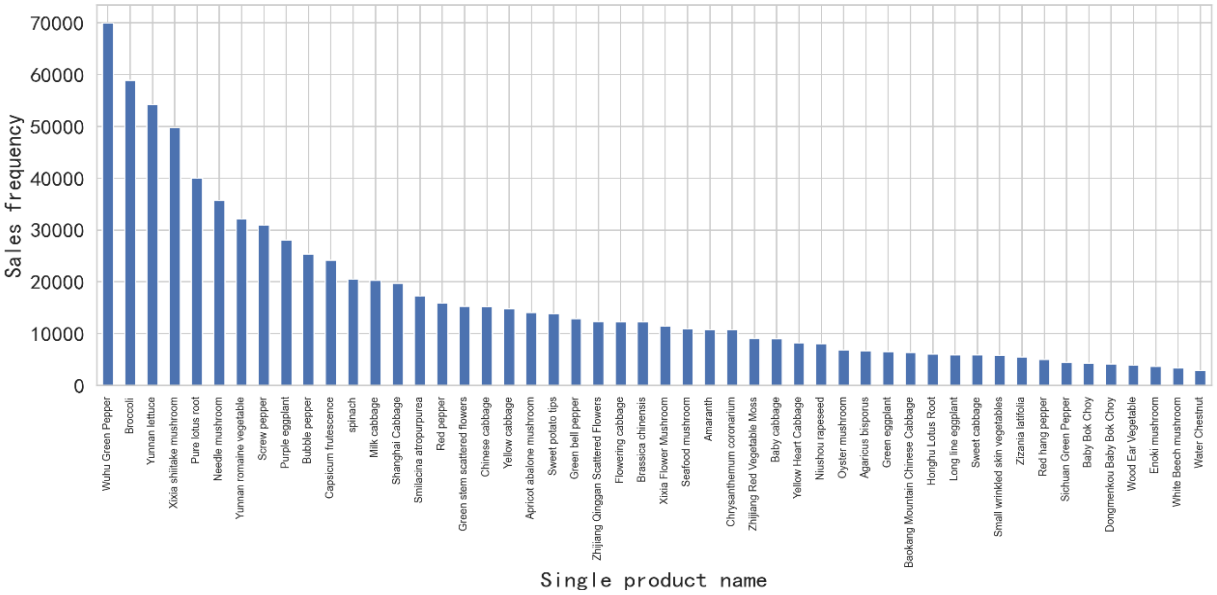

Figure 3. Single product sales category

Through the sales frequency bar chart, we can see that Wuhu Green Pepper is the first type of household dish, Broccoli, Yunnan lettuce, Xixia shiitake mushroom is the second type of daily side dishes, and the rest are the third type of other dishes. According to the sum of each vegetable category, the sales frequency of each category from large to small is Flower leaf class (331969), Capsicum (207996), Edible fungi (148424), Cauliflower category (86570), Aquatic rhizomes (58647), Solanaceae (44898).

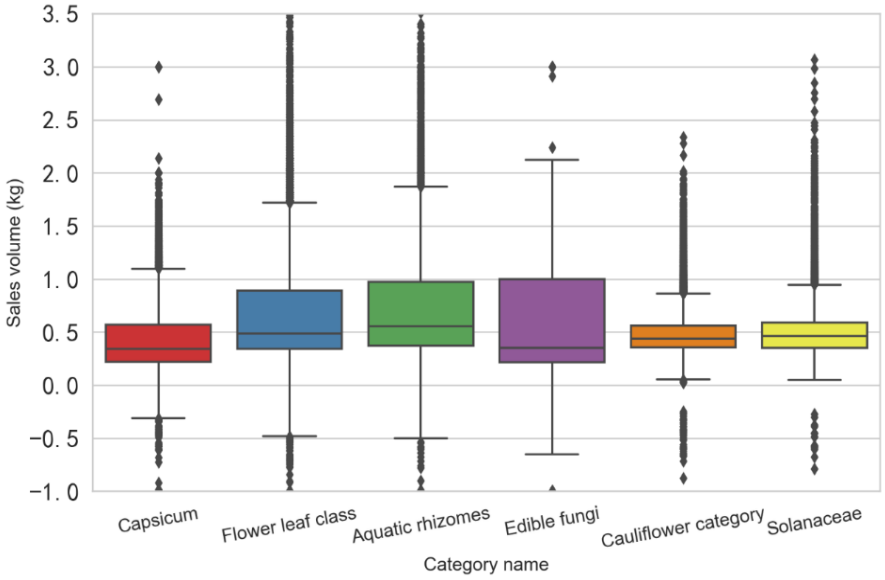

Figure 4. Box chart of sales volume of different categories

Using Jupyter and data-driven methods, the box plot distribution of purchase quantities for various vegetable categories was obtained as shown in Figure 4. By further specifying the driving force, we can obtain the sales situation box diagram of Edible Fungi as shown in Figure 5. The following text is presented in the Edible Fungi category.

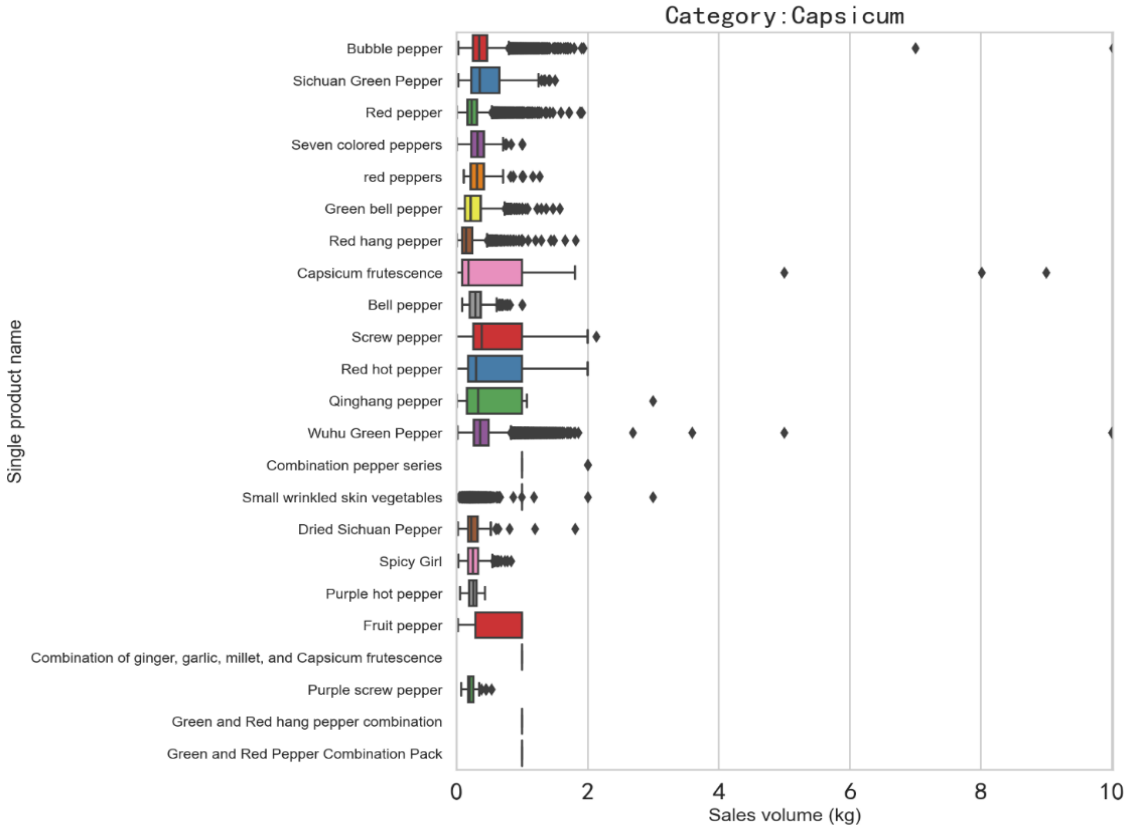

Figure 5. Edible Fungi's sales box chart

3. Supermarket Sales Strategies Integrating Statistics and Forecasting

In the process of supermarket sales forecast, we assume that the category is the unit of the replenishment plan, with the daily replenishment volume and pricing from July 1-7, 2023, as the range of prediction, to obtain the wholesale price of each item. Based on this, the cost, price, sales

volume, and cost-profit rate of each item are constructed. Then, the weighted wholesale price, weighted sales price, and weighted cost profit margin of each category can be obtained through the weighted idea as shown in Table 1 [8].

Table 1. Weighted wholesale prices, weighted sales prices, and weighted cost-profit margins for each category

	Classification Name	Sale Date	Total Sales Volume	Total Profit	Weighted Wholesale Price	Weighted Sales Price	Weighted Cost Profit Margin
0	Aquatic rhizomes	2020-07-01	4.850	25.49812	9.234161	14.491505	0.706978
1	Aquatic rhizomes	2020-07-02	4.600	20.48956	7.093748	11.548000	0.710128
…	…	…	…	…	…	…	…
6471	Solanaceae	2023-06-29	11.511	33.90844	4.555986	7.501729	0.632503
6472	Solanaceae	2023-06-30	24.530	83.61505	4.696215	8.104900	0.729765

Then, according to the classification name, the corresponding sales data are selected from the original data, including total sales volume, total profit, wholesale price, sales price, cost profit margin, and so on. The correlation coefficient matrix is visualized using a thermal map, where the thermal map uses color to represent the strength of the correlation, light color for high correlation and dark color for low correlation. As shown in Figure 6.

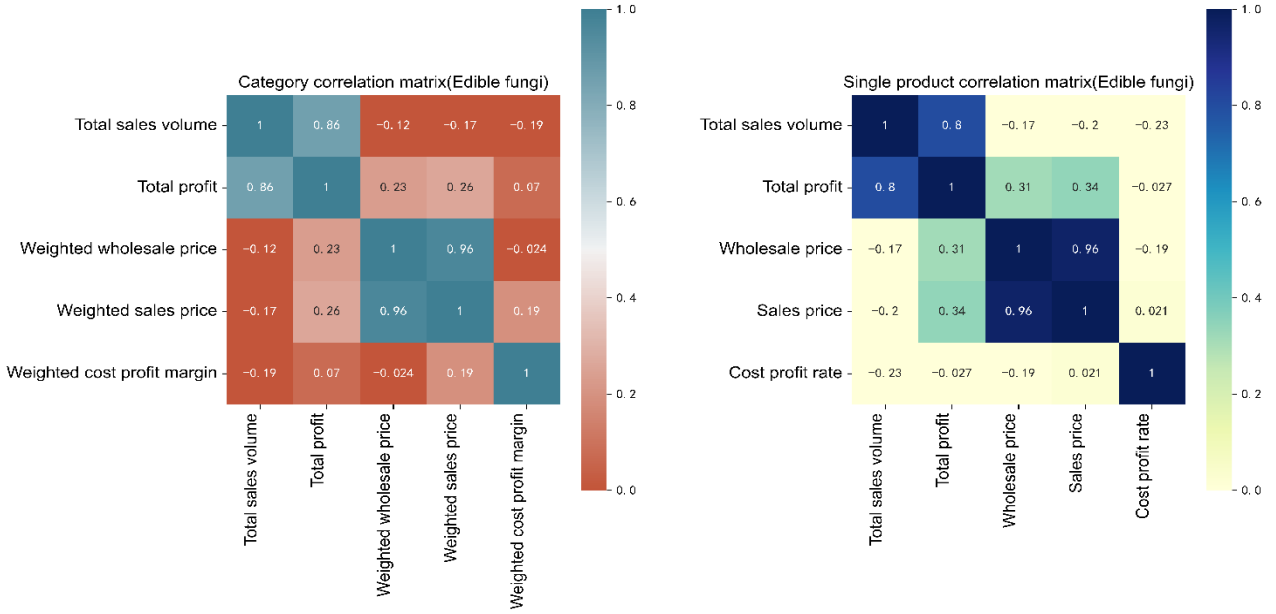

Figure 6. Thermal map of profit sales correlation coefficient

Set the weighted cost profit rate under different categories in the sales data as X and other variables as Y. By calculating the Spearman correlation coefficient ρ, we can obtain the class-weighted cost profit rate and the single product cost profit rate. In the Spearman test: The correlation coefficient between the Cauliflower category and weighted selling price is 0.98, and the correlation coefficient between Solanaceae and selling price is 0.91, showing a strong positive correlation.

We set total sales as the independent variable x and weighted sales price as the dependent variable y, and add a constant term ε, obtain a univariate linear regression model with $y = \beta_0 + \beta_1 x + \varepsilon$, and plot the slope, intercept, correlation coefficient R, P value, and standard error obtained from the linear regression analysis. Draw the distribution map of Edible fungi's total sales volume and the scatter plot of total sales volume and weighted sales price as shown in Figure 7 [9].

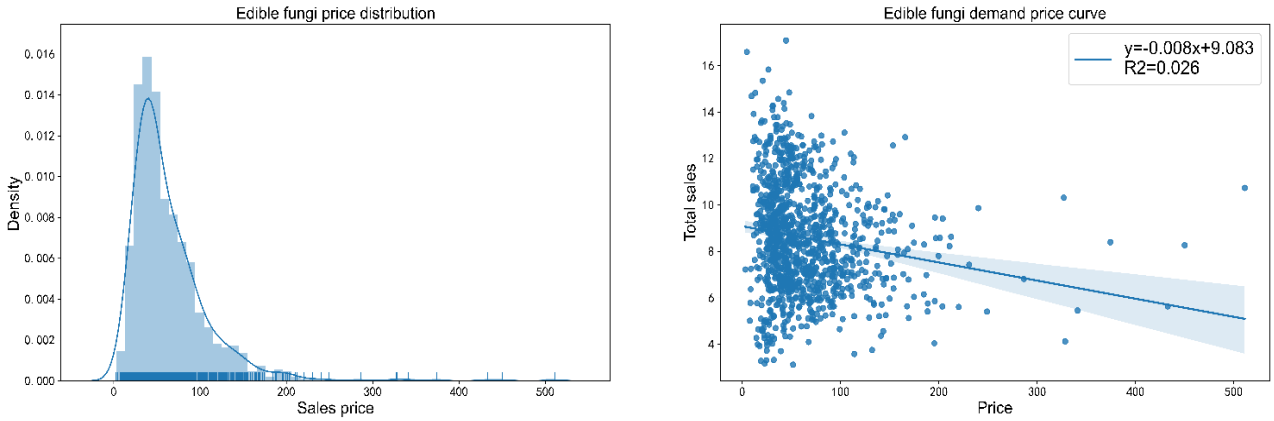

Figure 7. Edible fungi's total sales volume distribution chart and scatter plot of total sales volume and weighted sales price. The left column includes histograms and kernel density estimators. The right column includes a scatter plot that fits the regression line and displays the parameters and R-squared values of the regression equation.

Additionally, based on the given sales volume data, we use the ARIMA model for forecasting. By observing the data's expectation and correlation coefficients, we conduct a weak stationarity test and use the ADF test statistic. Then, using AIC, we determine the AR parameters and construct the autocorrelation coefficient plot and partial autocorrelation coefficient plot for Edible fungi as shown in Figure 8. We obtain various statistical indicators as shown in Table 2 [10].

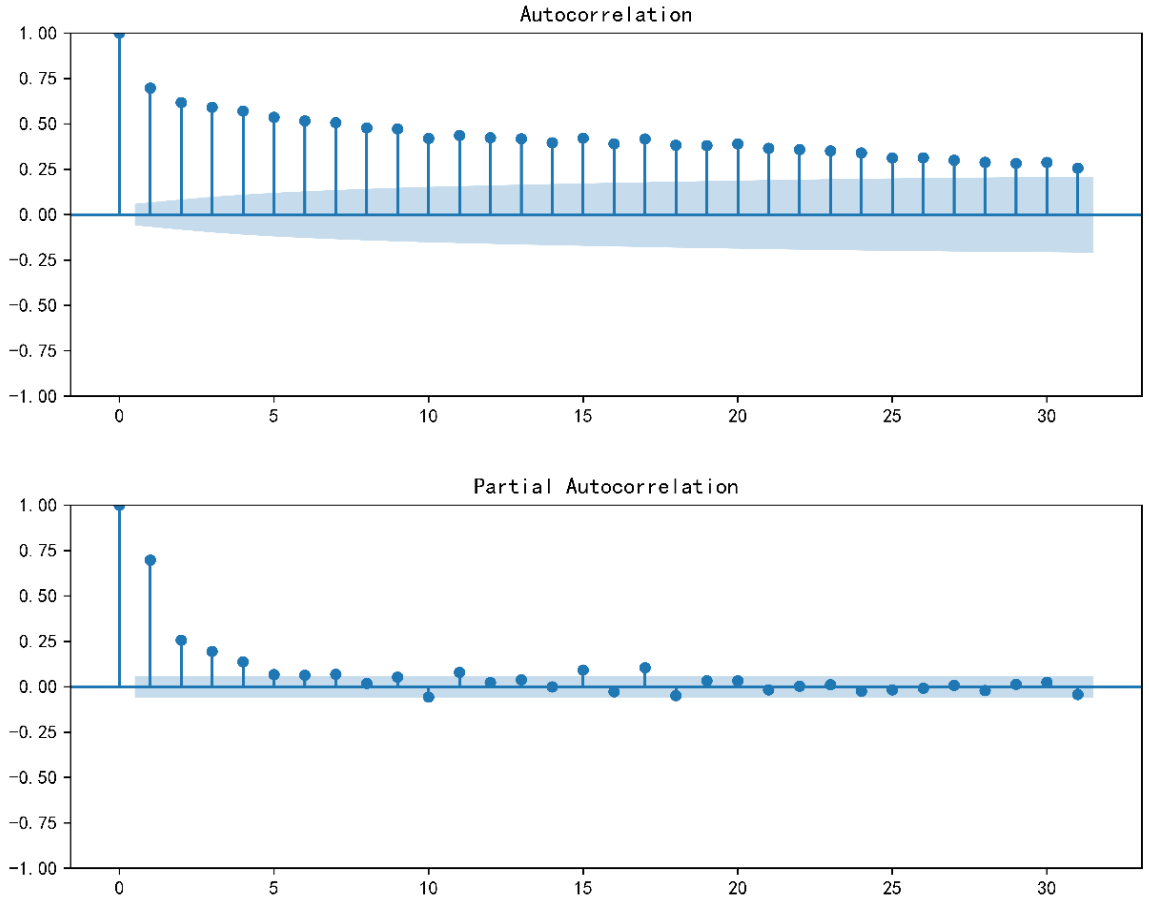

Figure 8. Autocorrelation coefficient plot and partial autocorrelation coefficient plot of Edible fungi

Table 2. Data from various statistical indicators

	Aquatic rhizomes	Flower leaf class	Edible fungi	Cauliflower category	Capsicum	Aquatic rhizomes
Test Statistic	-2.337969	-3.280195	-3.574731	-4.487570	-2.931902	-2.337969
AR Parameters	0.160014	0.015787	0.006263	0.000207	0.041762	0.160014
Optimal parameter	(2,1,1)	(1,1,1)	(2,1,3)	(3,1,2)	(1,1,0)	(2,1,1)

Finally, the model is used to predict the maximum returns from July 1 to 7, 2023, and the forecast results shown in Table 3 are obtained.

Table 3. Maximum profit forecast for each category

	Aquatic rhizomes	Flower leaf class	Edible fungi	Cauliflower category	Capsicum	Aquatic rhizomes
2023-7-1	125.9	844.4	276.8	285.6	252.1	241.6
2023-7-2	127.1	850.0	285.1	286.7	252.3	240.7
2023-7-3	97.5	976.8	249.8	186.1	215.4	182.6
2023-7-4	97.5	976.8	262.4	184.6	215.4	182.6
2023-7-5	105.6	976.8	252.5	184.8	215.4	182.5
2023-7-6	105.6	976.8	270.5	184.7	215.4	182.5
2023-7-7	103.0	976.8	256.7	186.1	215.4	182.5

Based on the prediction results, we can see that the profit forecasts for each category fluctuate roughly between 100-125 (Aquatic rhizames), 840-850 (Flower leaf class), 250-270 (Edible fungi), 185-190 (Cauliflower category), and 215-220 (Capsicum). And the returns of Aquatic rhizames and Flower leaf class show a relatively stable growth trend, while the returns of Edible fungi, Cauliflower category, and Capsicum fluctuate greatly, and the growth trend is not very stable.

4. Conclusions

This article first uses a time-based statistical analysis model to visualize and analyze the sales frequency of vegetable categories in a fresh supermarket. It concludes that Wuhu green peppers are the most popular household category, followed by broccoli, Yunnan lettuce, and Xixia shiitake mushrooms, which are commonly used as side dishes. Other vegetable categories have relatively low distribution. By observing the data fluctuations of different categories' sales at different times, it can be seen that the sales of household and side dish categories generally fluctuate before dinner time.

Secondly, based on the visualization-driven approach, the article calculates the Spearman rank correlation coefficient between sales volume and total sales, total profit, wholesale price, and selling price. It determines the relationship between sales volume and these factors, and obtains the slope, intercept, correlation coefficient R, P-value, and standard error. Then, a simple linear regression model is constructed to determine that the higher the cost-profit ratio of each category, the higher the cost and selling price, and the lower the total sales.

Finally, based on the parameters of the model, demand price curves for different categories are built, and the maximum revenue for each category from July 1st to 7th in 2023 is predicted. Based on the analysis results of this article, important references can be provided for replenishment and pricing decisions in future supermarkets.

References

[1] Chen Long. Empirical study on disposable income and consumption expenditure of urban residents: based on a univariate linear regression model [J]. China New Communications, 2019,21 (01): 233-234

[2] Zhao Jiabao, Chen Jie, And Xia, et al. Prediction and Analysis of Grape Yield in Turpan City Based on ARIMA Model [J]. Jiangsu Science and Technology Information, 2019, 36 (31): 34-39

[3] Wang Hui. Research on Ramie Fiber Yield Prediction Based on ARIMA Model [J]. Journal of Jiangsu Vocational and Technical College of Economics and Trade, 2023, (05): 8-10.

[4] Li Yuan, Liu Yutian, Feng Liwei. Feature extraction and fault detection of nonlinear dynamic processes based on Spearman correlation analysis [J]. Journal of Shandong University of Science and Technology (Natural Science Edition), 2023,42 (02): 98-107.

[5] Jia Ke, Yang Zhe, Wei Chao, et al. Pilot protection of new energy transmission lines based on Spearman level correlation coefficient [J]. Power System Automation, 2020,44 (15): 103-111

[6] Han Meng. Analysis of Consumer Price Index and Retail Price Index of Goods Based on Univariate Linear Regression Analysis [J]. Modern Business, 2020, (17): 12-13.

[7] Ren Jianying. Univariate linear regression analysis and its application [J]. Cai Zhi, 2012, (22): 116-117

[8] Hao Jinhong Exploration of the Application of Weighted Thinking in Data Analysis [J]. Mathematical Bulletin, 2019,58 (03): 29-32

[9] Wu Xinlin, Chen Shiyu. Application of Univariate Linear Regression Model in Educational Economy Prediction [J]. Journal of Hubei Second Normal University, 2023,40 (08): 25-30

[10] Ding Jing, Huang Xiaoming, Yu Chenglong. Power material prediction based on multi factor fusion and ARIMA [J]. Energy and Environmental Protection, 2022, 44 (07): 227-231.