Journal of Innovations in Economics & Management (2024)

Journal of
Innovations in Economics & Management
journal homepage: www.journal-iem.online

Paper Type: Original Article

Research on the distribution patterns and interrelationships of sales volume of various vegetable categories and individual products

Hao Li[1, #] Siyi Huang[2, #] Rui Hu[3, #] Minghao Liu[4, #] Junhao Li[5, #] Fuqiang Huang[6, #]

1. College of Fire Protection Engineering, China People's Police University, Hebei, China

2. School of Mathematics and Statistics, Jishou University, Jishou, China

3. School of economics and management, Guangxi Normal University, Guilin, China

4. College of Water&Architectural Engineering, Shihezi University, Shihezi, China

5. First Clinical Medical College，Southern Medical University，Guangzhou, China

6. School of Electrical and Control Engineering, Shenyang Jianzhu University, Shenyang, China

Co-first author

Abstract

This study conducts a comprehensive analysis of vegetable sales data. Using Python, the collected data was integrated, categorized, and checked for missing values with the "missingno" library. Anomalies were detected using box plots, revealing a few outliers, which were retained as they more accurately reflect real business phenomena. The analysis explored distribution patterns and interrelationships of sales quantities across different vegetable categories and individual items. Time series decomposition revealed significant seasonal variations in sales volumes throughout the year. Model accuracy was validated through residual analysis, and missing values were imputed using the Prophet model. Dynamic Time Warping (DTW) calculated distance matrices between categories to uncover similarities. K-means clustering analyzed sales trends and seasonal patterns of individual items, with DTW providing detailed similarity analysis within clusters. This approach identified correlated sales trends, such as the high correlation between Chinese cabbage and bell peppers, indicating that consumers may prefer to purchase these vegetables together. These findings offer valuable insights for optimizing supermarket restocking strategies.

Keywords: Vegetable Sales Data, Python, Missing Values, Anomaly Detection, Box Plots, Time Series Decomposition, Prophet Model, Dynamic Time Warping (DTW), K-means Clustering, Sales Trends, Seasonal Patterns, Supermarket Restocking.

1 | Introduction

Understanding the sales dynamics of vegetables is crucial for optimizing inventory management and improving customer satisfaction in the retail sector. This study investigates the distribution patterns and interrelationships of vegetable sales data, utilizing advanced data analysis techniques to reveal valuable insights.

By integrating and analyzing vegetable sales data through Python-based methods, including anomaly detection, time series decomposition, and clustering, we aim to identify significant trends and correlations. These insights are intended to support more effective supermarket restocking strategies and enhance overall inventory management.

2 | Data preprocessing

In order to better observe the structure and characteristics of the dataset, we first perform data type statistics on it. This not only helps us understand the characteristics of data, but also provides direction for subsequent data processing and analysis. The solution results are shown in Table 2, Sales Date: This is a date type that can be used for time series analysis; Scan code sales time: This is an object (string) type that represents the specific time when the code was scanned. If more accurate time analysis is needed, it may be necessary to convert it into a time type; Single item code and classification code: Both are integer types and are commonly used as identification codes; Sales volume (kg), sales unit price (yuan/kg), wholesale price (yuan/kg), and single item loss rate (%): These are all floating-point types, representing the sales quantity, selling price, wholesale cost, and loss situation of the product, respectively; Sales type, discount sales, item name, and category name: These are all object (string) types that represent various descriptive information.

2.1 | Missing value statistics

Using the "matrix" method of the "missingno" library, we generated a missing value matrix graph to visually demonstrate the integrity of the data. In this matrix, white represents missing values, while black represents the presence of data. Through this method, we can quickly understand which columns have missing values and their distribution. By observing Figure 1, we found that there are no missing values in the dataset.

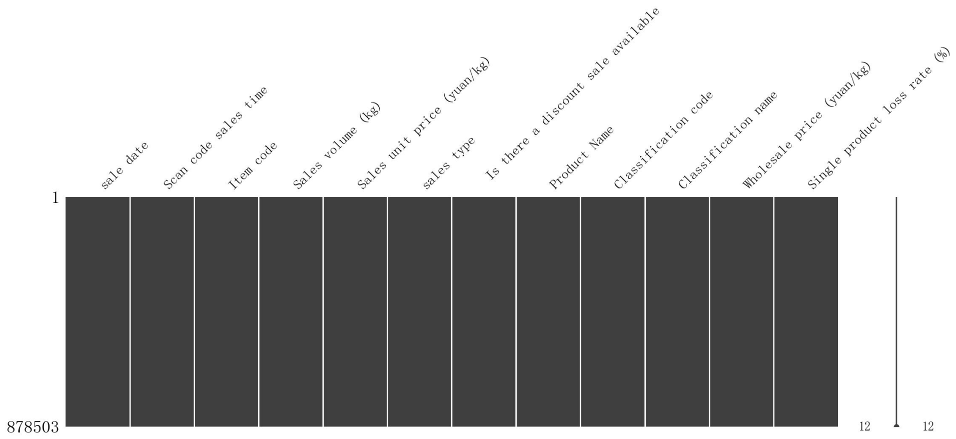


Figure 1 Missing value statistics

2.2 | Outlier statistics

In this article, we use box plots to detect outliers in the sales volume, selling unit price, wholesale price, and single item loss rate of vegetable sales data. This method can help us visually identify and analyze data points that deviate from the normal range, ensuring that our analysis is based on reliable data and providing insights for special situations or different sales strategies in the data.

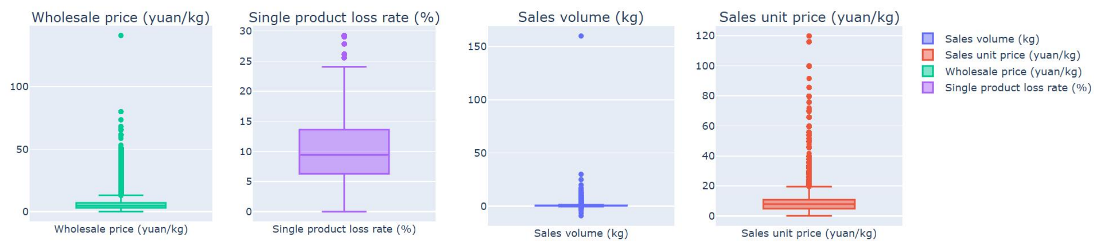


Figure 2 boxplot

According to the box diagram, we found that:

- Sales volume (kg): Although the sales of most dishes are concentrated in a relatively low range, there are some abnormally high sales points. For example, changes in demand before and after holidays or seasonal changes may lead to a sudden increase in sales of certain dishes.

- Sales unit price (yuan/kilogram): The sales unit price of most dishes is relatively stable, but the unit price of some dishes is abnormally high. This is a reasonable existence in real life, such as price increases caused by shortages in certain vegetable supplies.

- Wholesale price (yuan/kilogram): The abnormal value of wholesale price is similar to the sales unit price. The wholesale prices of certain dishes are abnormally high, possibly due to a shortage of supply in the market.

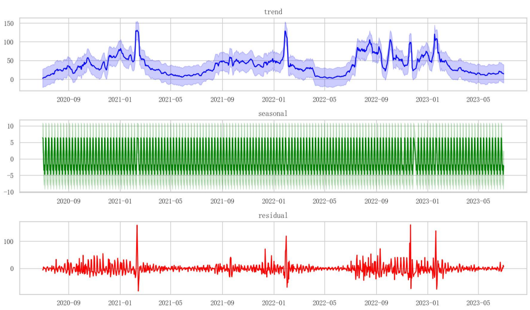
- Single item loss rate (%): Although the loss rate of most dishes remains at a low level, there are also some dishes with abnormally high loss rates. This may be due to certain dishes being more susceptible to damage.

In summary, for real business data, outliers may represent real and important business phenomena. Removing these outliers may make the analysis results smoother and more consistent, but this may overlook or misunderstand the true business dynamics. Therefore, this article does not recommend deleting these outliers. Instead, it is more important to conduct in-depth exploration and explanation of these outliers. This can help us better understand the real situation of the business and provide more valuable insights for decision-making.

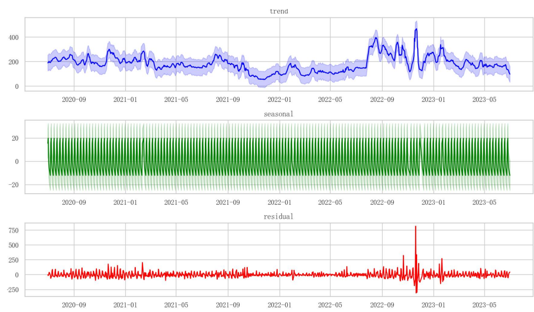
3 | Distribution pattern and correlation analysis of categories

3.1 | Analysis of Distribution Patterns

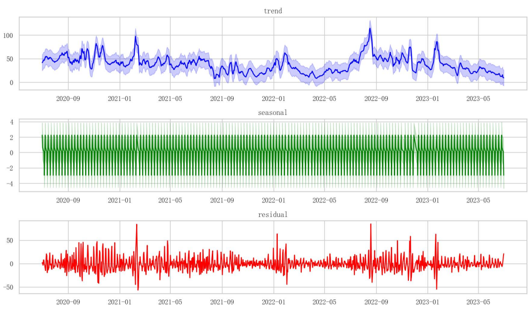
In order to solve the distribution pattern of various categories of vegetable products, we choose to use time series decomposition method. This method can break down sales data into three parts: trends, seasonality, and residuals. Through this decomposition, we can clearly see the long-term trend of product sales, identify cyclical patterns related to time changes, and explore any abnormal sales behavior. This method is particularly suitable for vegetable sales as it is influenced by various factors such as season, weather, and holidays. Time series decomposition provides us with a tool to gain a deeper understanding of these complex patterns, thereby providing powerful decision support for businesses[1],[2]. The result of using Python to solve is shown in Figure 3.



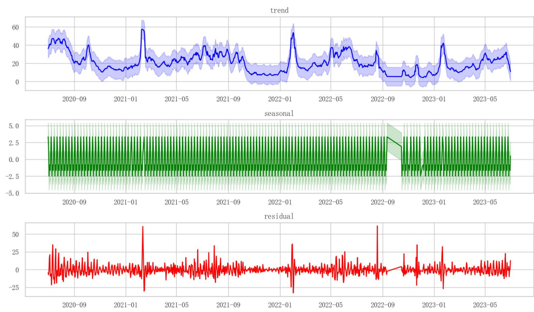
Aquatic rhizomes



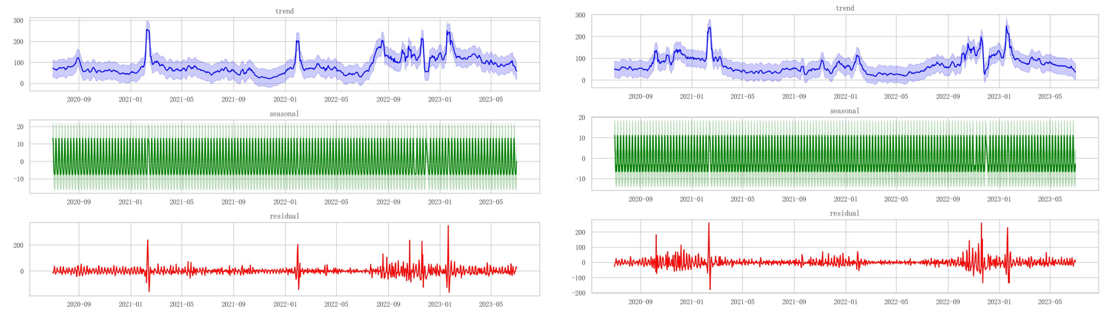
Flowers and leaves



florescent vegetables



Eggplant type



Chili peppers **edible fungi**

Figure 3 Sales data time series

From the above figure, we can see that the residual plot shows that most of the deviations are concentrated around 0, indicating that the model has relatively successfully captured the main patterns in the data, including seasonality and trends. The time series decomposition of six categories of vegetables shows clear trends and seasonal variations. From the seasonal chart, we can observe that the sales of each vegetable category are generally higher on weekends, which may be due to people having more time for shopping and cooking on weekends.

From the trend chart, we can observe that the sales of vegetables in each category are significantly higher from September to January each year than in other months. This may be related to the arrival of winter and the shopping frenzy before the Spring Festival. During the Spring Festival, families usually hold grand dinners, which involve a lot of cooking and require more dishes than usual.

These observation results provide us with a deep understanding of the sales of various vegetable categories, and provide a basis for businesses to further develop sales strategies and promotional activities.

3.2 | Correlation analysis

Dynamic Time Warping (DTW) is a popular technique commonly used for pattern matching and similarity measurement in time series data. The main advantage of this method is that it can compare two time series, even if their lengths are different or their patterns have temporal offsets. It aligns two sequences as closely as possible by bending the timeline, allowing for comparison between the two sequences[3],[4].

In this article, we have sales data for different categories of vegetables. In order to understand the sales dynamics between these categories, we have decided to use DTW. For this, we calculated the DTW distance between all possible category pairs[5],[6]. In this way, we can obtain a distance matrix where each element represents a measure of sales pattern similarity between two categories.

Through this method, we can identify which vegetable categories have similar sales patterns, which may indicate that they are influenced by similar market factors or have similar popularity among shoppers[7],[8]. In addition, this analysis can also help us discover which categories have significantly different sales models from other categories, which can provide valuable insights for market strategies.

However, due to the previous time series decomposition automatically estimating and filling in these missing values, incomplete data was not detected. Therefore, when we perform DTW, it will display the presence of missing values in the data. In order to find missing values, we drew the original sales data graph as shown in Figure 4, and found that "cauliflower category" had a missing value on the date "2022-03-29". The term 'eggplant' has missing values on multiple dates, specifically from '2022-09-12' to '2022-10-14' and '2022-12-05'.

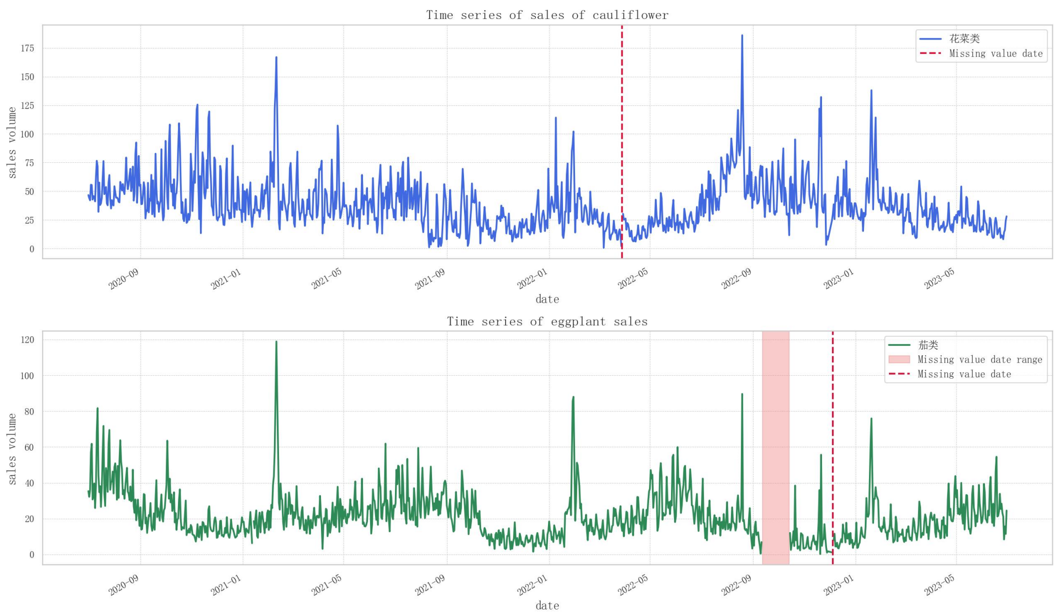


Figure 4 Missing value display

In the previous time series decomposition analysis, we found clear trends, seasonality, and other periodic changes in the data, but traditional imputation methods such as linear interpolation or using before and after averages may not take into account these complexities in the data.

Facebook Prophet is a tool designed specifically for time series forecasting, which can automatically detect trends and seasonality in time series data and make predictions based on them. After decomposing the time series data, we found significant seasonality and trends in the sales data of categories such as "cauliflower" and "eggplant". This means that simply filling in missing values may lead to analysis bias.

Therefore, we choose to use Prophet to fill in missing values. It can provide a more reasonable prediction value for missing dates based on the trends and seasonality captured by existing data. This not only ensures that the filled data is statistically consistent with the original data, but also makes the predictions on these dates more reliable and accurate. The fitting effect is shown in Figure 5, and it can be seen that Prophet has excellent performance[9],[10].

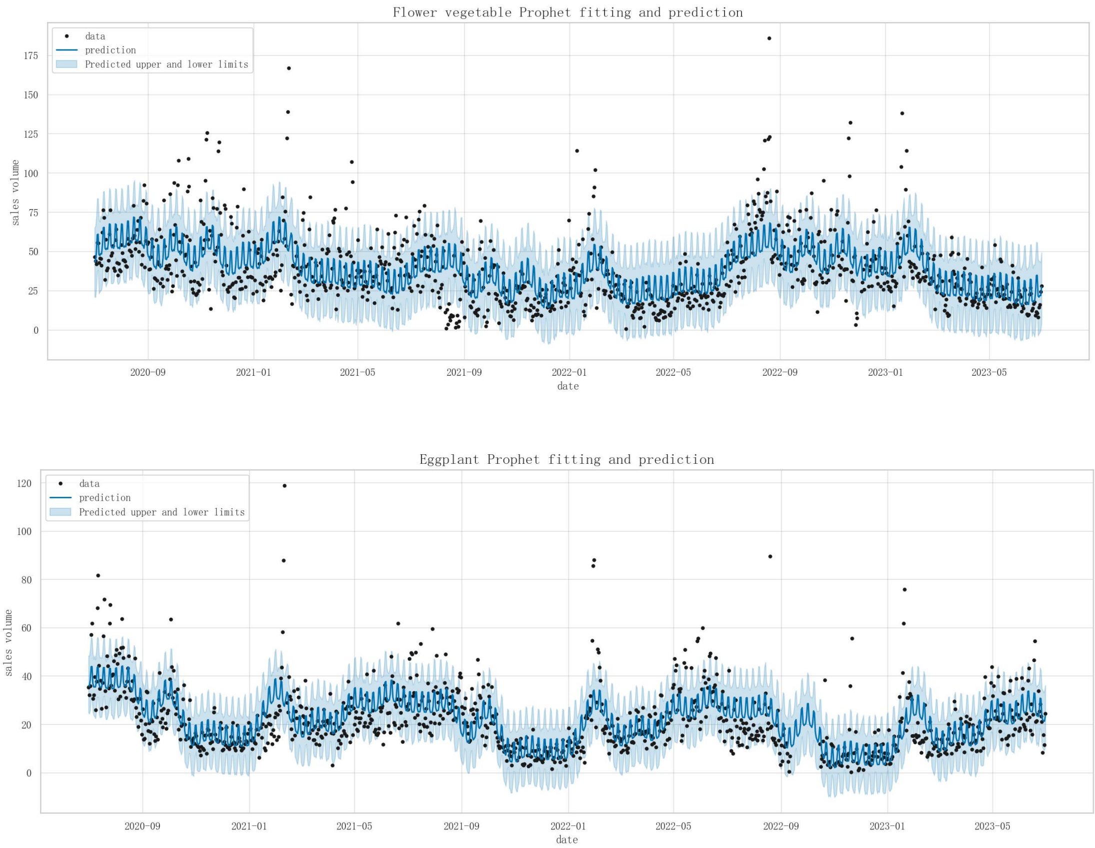


Figure 5 Fitting effect diagram

We use the complete data after filling to solve the DTW distance between category pairs, such as "eggplant" and "cauliflower". To compare these two time series, we use a distance matrix D with a size of $m \times n$, where m is the length of the "eggplant" data and n is the length of the "cauliflower" data. Each element $D(i,j)$ represents the distance between the i-th point of the "eggplant" time series and the j-th point of the "cauliflower" time series. Our goal is to find a path from D (1,1) to $D(m,n)$ that minimizes the sum of distances on this path. The specific calculation formula is:

$$D(i,j) = d(i,j) + \min\{D(i-1,j), D(i,j-1), D(i-1,j-1)\} \quad (1)$$

Among them, $d(i,j)$ is the Euclidean distance between two points. Finally, $D(m,n)$ will provide us with the DTW distance between two time series. For all our vegetable categories, we can calculate the DTW distance between each pair of categories by repeating the above process, thus obtaining a distance matrix representing the similarity of sales patterns between the two categories[11],[12].

The result of using Python to solve is shown in Figure 6.

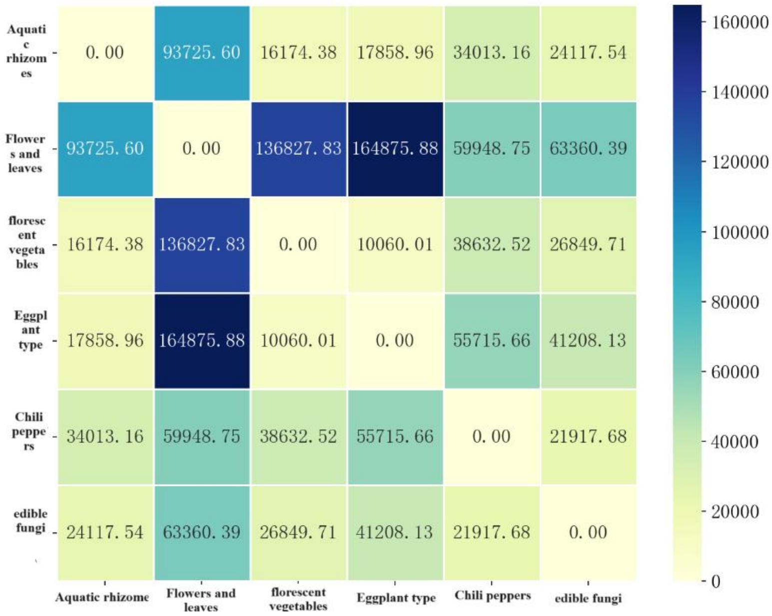


Figure 6 DTW distance between categories

Based on the results of Dynamic Time Warping (DTW) in Figure 6, we can delve into the interrelationships between vegetable varieties as follows:

(1) The DTW distance between "cauliflower" and "eggplant" is the smallest, indicating that their sales models are very similar. This may mean that consumers' decisions to purchase these two vegetables are influenced by similar external factors, such as seasonal changes, weather conditions, or promotional activities. Furthermore, this may also imply that when a merchant promotes or discounts one type of vegetable, the sales of the other vegetable may also be affected.

(2) The DTW distance between "eggplant" and "flower leaf" and "aquatic rhizome" and "flower leaf" varieties is relatively large, indicating that their sales models have significant differences. This may mean that they are complementary or alternative in consumers' shopping baskets. For example, there is a significant difference in sales models between "eggplants" and "flowers and leaves". When the price of "flowers and leaves" increases, consumers may switch to buying "eggplants", and vice versa.

(3) Other variety combinations: Their DTW distance is in a moderate range, indicating that there are both similarities and differences in sales patterns among these varieties. For example, "chili peppers" and "edible mushrooms" may have some consumers who like chili stir fried mushrooms, while others do not like chili peppers and mushrooms made together.

4 | Distribution pattern and correlation analysis of individual

products

4.1| Analysis of Distribution Patterns

To gain a deeper understanding of the sales distribution of each individual product, we can also use time series decomposition to observe the sales patterns of individual products. Similar to the previous analysis of vegetable categories, we can separately examine the trends, seasonality, and residuals of each individual product. We can see from the data that there are a total of 251 items, and only the time series decomposition results of 6 items are shown below, as shown in Figure 7.

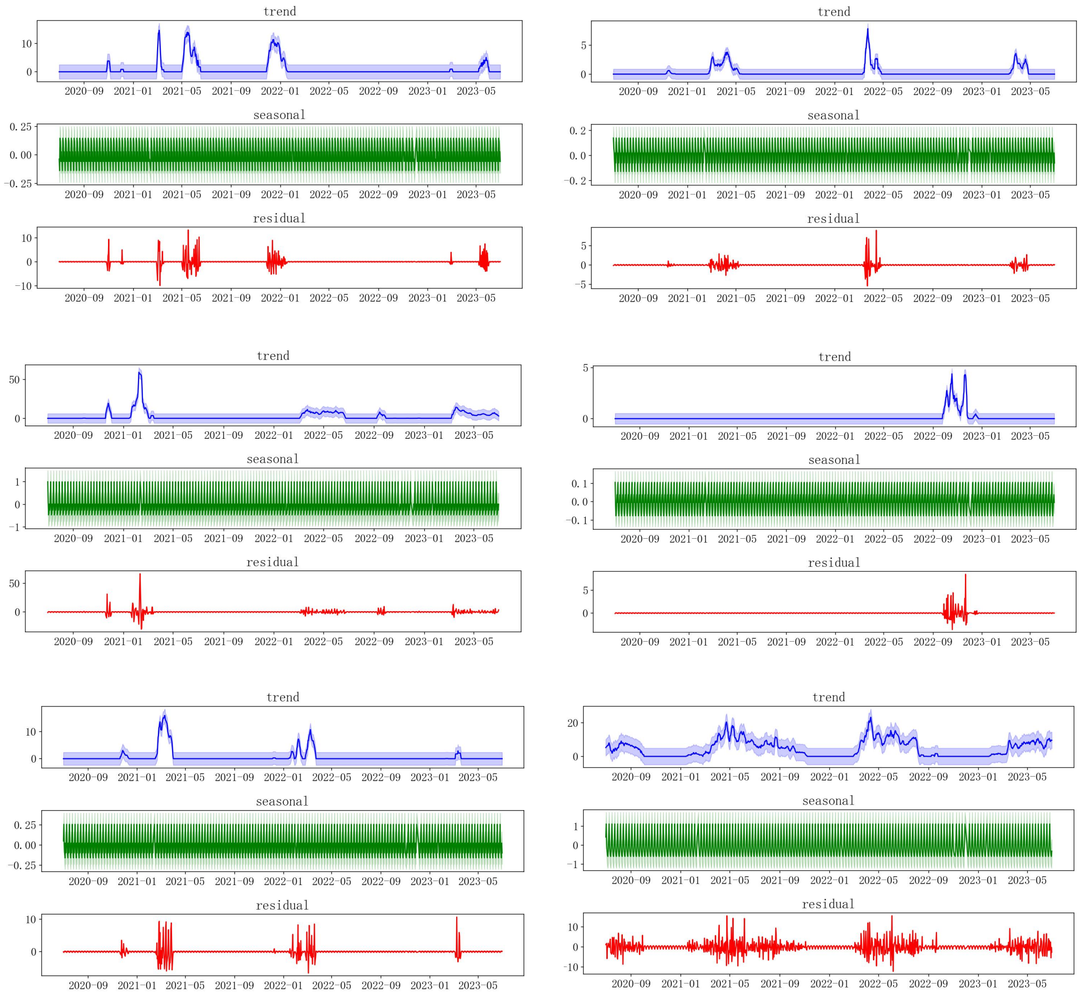


Figure 7 Sales data time series

From the above figure, we can see that the residual plot shows that most of the deviations are concentrated around 0, indicating that the model has relatively successfully captured the main patterns in the data, including seasonality and trends. The time series decomposition of each individual item displays clear trends and seasonal variations. From the seasonal chart, we can observe that the sales of each vegetable are generally higher on weekends, which may be due to people having more time for shopping and cooking on weekends. This is similar to the previous category analysis results. The trend chart shows that certain products have clear sales peaks or valleys in certain months of the year, such as chrysanthemum, which has sales from September to April of the following year. The sales in other months are basically zero, and the sales in October and February of the following year are often the highest, which is closely related to local dietary habits and the maturity time of chrysanthemum. The sales distribution of other individual products also follows a similar pattern.

4.2 | Correlation analysis

Due to the wide variety of individual products and their unique sales trends, we need to delve into the interrelationships between each product. In order to more accurately identify strong correlations between individual products, we have decided to cluster the trend and seasonal components. By conducting detailed clustering analysis on these components, we can classify individual items into different categories. Under this classification, each item in the category will exhibit similar sales patterns. For example, a certain type of product may exhibit a steady upward sales trend, accompanied by significant weekly sales cycle fluctuations. And another type of product may show a gradually declining sales trend, with specific sales peaks every year. Through this classification method, we can have a more systematic understanding of the sales characteristics of products.

In order to further explore the sales similarity between individual products, we chose to use the Dynamic Time Warping (DTW) method instead of the traditional Pearson correlation analysis. DTW can more accurately capture patterns and trend similarities between two time series, especially when they may have temporal offsets. Next, for each clustering category, we use DTW to calculate the sales similarity between individual products within the category. In this way, we can not only identify which product sales trends are interrelated, but also gain a deeper understanding of the degree and characteristics of this correlation.

4.2.1 | K-means clustering

We have decomposed the sales data of each item into three parts: trend, seasonality, and residual, using the method of time series decomposition. Among them, trends represent the long-term sales patterns of goods, while seasonality reflects the sales fluctuations of goods within a certain period. By conducting cluster analysis on these two components, we hope to find products with similar sales patterns.

When using algorithms involving distance calculation, such as k-means clustering or k-NN, large-scale features may dominate the results, causing small-scale features to be ignored. Standardization can ensure that all features are on the same scale, thereby avoiding this bias. We first standardize the data[13].

$$Z_i = \frac{X_i - \mu}{\sigma} \tag{1}$$

To determine the optimal number of clusters, we used the elbow method. The core idea of the elbow method is that as the number of clusters increases, the sum of squared errors (WCSS) within each cluster will gradually decrease. When the increased number of clusters no longer significantly reduces WCSS, this point is called the "elbow" and is also the optimal number of clusters we choose. As shown in Figure 8.

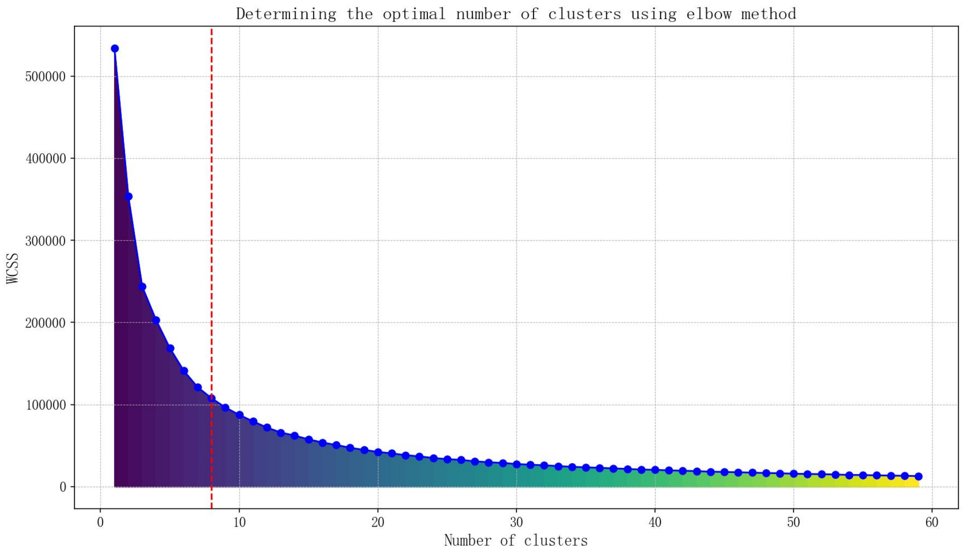


Figure 8 Determining the optimal number of clusters using elbow method

Looking for the elbow point, we have determined that the optimal number of clusters is 8. In these 8 categories, each item exhibits a similar sales pattern. For example, we found that a certain type of vegetable showed a stable upward trend in sales, accompanied by weekly sales cycle fluctuations. And another type of product shows a declining sales trend year by year, but there will be significant sales growth in specific seasons (such as holidays). The clustering results are shown in Figure 9.

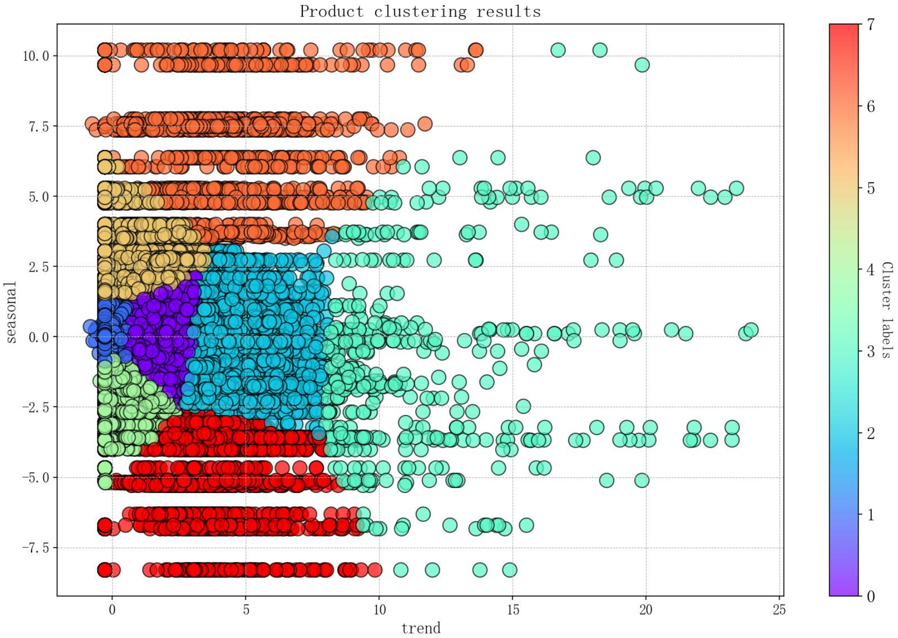


Figure 9 Product clustering results

4.2.2 | Dynamic Time Warping (DTW)

Due to the large number of categories in clustering, we take Cluster 1 as an example to conduct DTW research. Firstly, we examine the cluster centers of each cluster, as shown in Table 1.

Table 1 Cluster Center

trend	seasonal
0.254325	0.002770
10.697518	-0.060178
29.302580	-0.209747
1.966909	-0.905304
3.505316	1.371831
23.469085	-2.626900
25.039703	3.546196
73.703120	-0.541578

The centers of Cluster 1 are Trend=0.253564 and Seasonal=0.002091. Compared to other clusters, we can see that:

- The trend value of Cluster 1 is relatively small, which means that the vegetable sales trend of Cluster 1 is relatively stable.
- The seasonal value of Cluster 1 is also very small, which means that its vegetable sales do not have significant seasonal fluctuations.

Then we also conducted descriptive statistics on each cluster, where the mean sales trend of cluster 1 was 0.254 and the standard deviation was 0.852. The specific differences are shown in Figure 10 and 11. Descriptive statistics also imply that the sales trend of cluster 1's products is relatively stable.

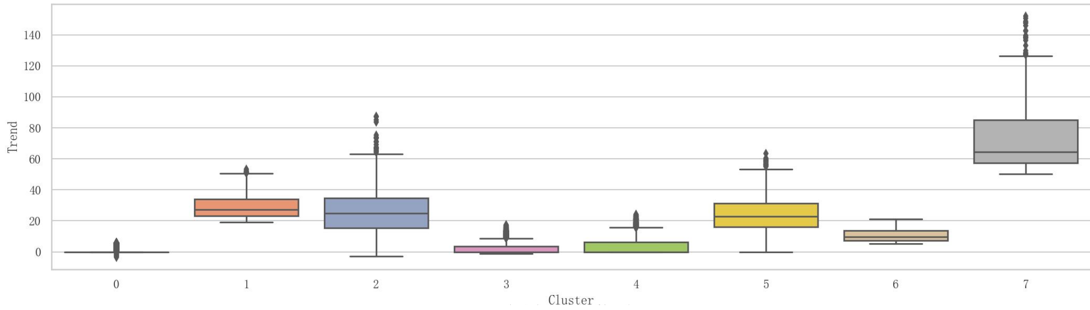


Figure 10 Trend distribution in each cluster

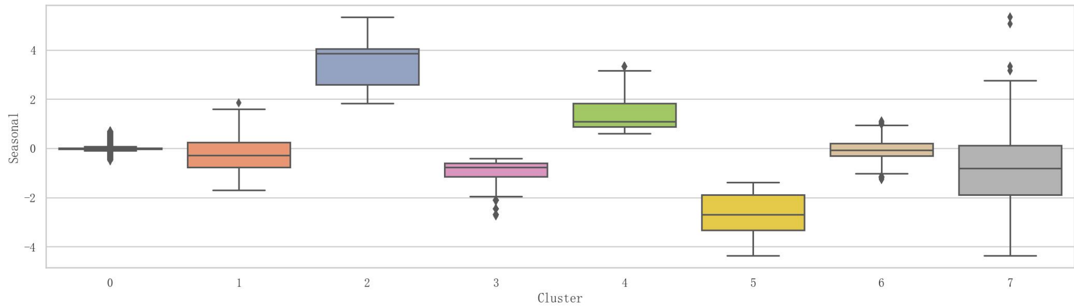


Figure 11 Seasonal distribution in each cluster

When we examine the individual product names in Cluster 1, we find that most of them are vegetables that we frequently consume throughout the year, which is consistent with the results of our analysis above.

Then we can further use DTW to analyze the similarity within Cluster 1. The calculation results are shown in Figure 12.

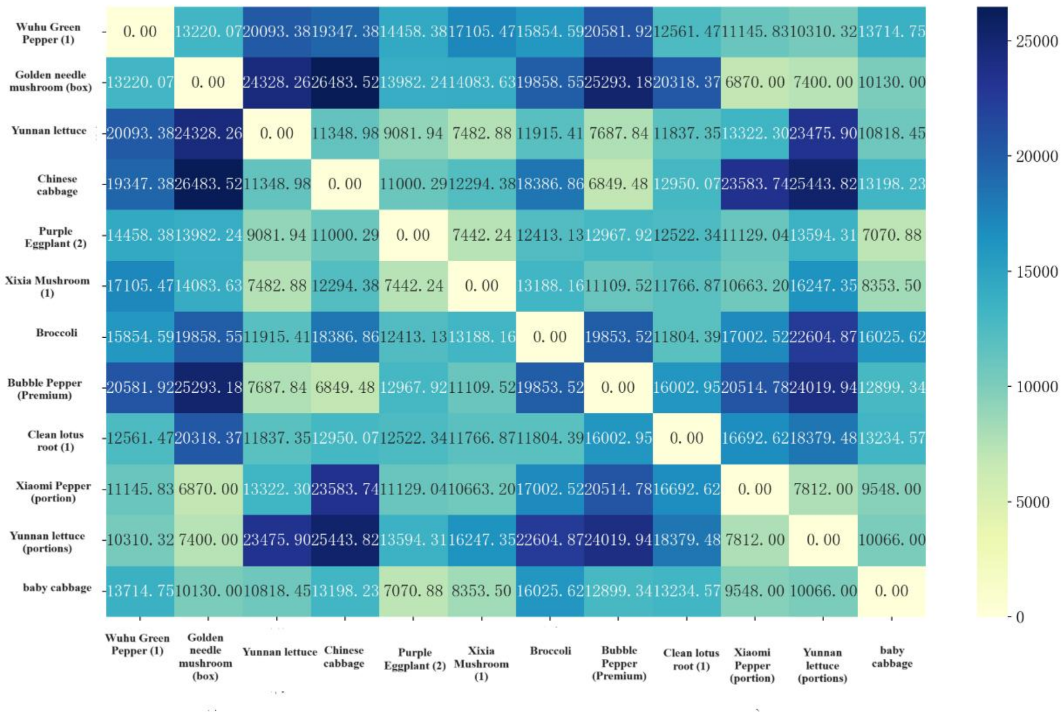


Figure 12 DTW distance of products in cluster 1

Based on the chart analysis, we can draw the following conclusion: Chinese cabbage and bubble pepper (premium) have the lowest DTW values. This not only means that their annual sales trends are relatively more stable compared to other products, but also that there is a high degree of correlation between them. Usually, consumers tend to purchase these two products together, which is in line with their real-life situations as they are often used together for cooking. Therefore, our relevant analysis results are reasonable.

On the other hand, the DTW values of shiitake mushrooms and Chinese cabbage (box) are the highest. This indicates that although their annual sales trends are similar, the degree of correlation between them is low and they are generally not likely to be purchased together.

5| Conclusion

This study effectively analyzed vegetable sales data through comprehensive data preprocessing, time series decomposition, and correlation analysis. We found that the dataset is well-organized, with no missing values and outliers that reflect real business dynamics rather than data errors. Time series decomposition revealed clear seasonal trends, with notable sales increases during weekends and from September to January.

Dynamic Time Warping (DTW) analysis highlighted significant correlations between vegetable categories, such as the similarity in sales patterns between "cauliflower" and "eggplant," suggesting that promotions for one may impact the other. K-means clustering further categorized products into distinct groups with similar sales patterns, aiding in understanding consumer behavior and optimizing inventory management. These findings offer practical insights for improving supermarket restocking strategies and overall operational efficiency.

References

[1] Sai H ,Ao L ,Dongyue Z , et al. Early warning of core network capacity in space-terrestrial integrated networks [J]. Journal of Systems Engineering and Electronics, 2024, 35 (04): 855-864.

[2] Chang Yanan, Duan Xingzhuo, Cui Jianqun, etc Opportunistic Network Routing Algorithm Integrating Unsupervised Learning Model X-Means [J/OL] Small microcomputer system, 1-13 [2022-08-26].

[3] Liu Yang, Zhou Haoyue, Lu Jinqi, etc Prediction of the incidence trend of influenza like cases in Jiaxing City, Zhejiang Province based on the Prophet model [J] Disease Surveillance, 2024, 39 (05): 629-633.

[4] Bi Zihang, Li Sumin, Zhang Longyu, etc Multi track time-series InSAR mining area 3D deformation monitoring and early warning combined with Prophet CNN model [J] Surveying and Mapping Bulletin, 2024, (05): 53-59 DOI:10.13474/j.cnki.11-2246.2024.0510.

[5] Zhang Shuhan, Cheng Yuehua, Jiang Bin A multivariate trend prediction method for solar cell arrays based on the STL Prophet Informer model [J] Space Control Technology and Applications, 2024, 50 (01): 35-45.

[6] Wei Meifang, Yang Jing, Huang Di, etc High loss line electricity theft detection method based on segmented dynamic time bending distance [J/OL] Southern Power Grid Technology, 1-9 [2022-08-26].

[7] Zhang Haiyan, Yan Wenjun, Zhang Limin, etc Research on Pilot Landing Skill Evaluation Based on Differential Thinking and Dynamic Time Warping (DTW) [J] Journal of Weapon Equipment Engineering, 2023, 44 (03): 124-130.

[8] Huang Zimeng, Yu Juan, Xiang Mingxu, etc PMU frequency anomaly detection and type recognition based on improved dynamic time bending [J] Power System Automation, 2022, 46 (24): 104-112.

[9] Ran Qisheng, Zhang Zhe, Han Jiexiang, etc Longitudinal protection scheme for DC distribution network lines based on improved dynamic time bending distance algorithm [J] Power Automation Equipment, 2022, 42 (12): 157-164.

[10] [10] Wen Hongbo, Liu Xianwei, Jiang Youxiang Reliability analysis of K-means clustering method in setting middle school entrance examination standards [J] Chinese Exam, 2024, (08): 69-78 DOI:10.19360/j.cnki.11-3303/g4.2024.08.008.

[11] Shi Jiangnan, Peng Changgen, Tan Weijie K-means++clustering method supporting differential privacy protection in Spark framework [J] Information Security Research, 2024, 10 (08): 712-718.

[12] Liu Jiahui, Zhang Ping, Cao Jinyin, etc Exploration of Disease Cost Clustering Analysis and Fine Management Based on K-means Algorithm [J] Health Economics Research, 2024, 41 (08): 37-40+44 DOI:10.14055/j.cnki.33-1056/f.2024.08.006.

[13] Li Maolin, Xiao Dongsheng A precise positioning method for personnel inside buildings using the fusion of the strongest base station and K-means [J/OL] Surveying and Mapping Science, 1-12 [2022-08-26].