Unveiling the Dynamics of Vegetable Category Sales Through a Data-Driven Approach

Sheng Wu[1,*], Jie Zheng[2], Xianlei Yuan[2], ChuYang Qiu[2]

[1]School of Information Engineering and Automation of Kunming University of Science and Technology, Kunming, China, 650500

[2]Information Engineering College of Kunming University of Arts and Sciences, Kunming, China, 650222

*Corresponding author: zjhunter_bachelor@163.com

Abstract. In the context of the development and application of big data technology, it is particularly important to use a data-driven approach to understand the changes in consumer demand for different vegetable categories and the relationship between them and sales volume to optimize supply chain management and formulate effective marketing strategies. In this paper, three machine learning algorithms, namely K-means cluster analysis, quadratic parabolic regression, and Pearson correlation analysis, were used to analyze the sales of vegetable categories, and the characteristics of people's purchase of vegetables and the correlation coefficients between sales volume and category were -0.12 and 1.00, indicating that the correlation between purchase characteristics and categories was weak, and the sales volume was completely dependent on the category of vegetables.

Keywords: K-means clustering, Parabolic regression, Correlation coefficients, Data - Driven.

1. Introduction

In today's rapidly changing market environment, a deep understanding of consumer demand for different vegetable categories and the complex relationship between them and sales volume is key to achieving supply chain optimization and developing effective marketing strategies. The rapid development of information technology, especially the wide application of big data technology, provides unprecedented data support and analysis tools for market analysis. Against this backdrop, adopting a data-driven approach to analytics can not only capture subtle changes in consumer behavior more accurately but also help companies extract valuable information from large amounts of sales data to gain an advantageous position in a highly competitive market. Therefore, the use of advanced data analysis technology to explore the relationship between vegetable categories and sales volume, and how to predict market trends and consumer needs through these relationships, has become an indispensable part of modern supply chain management and market strategy formulation.

Firstly, the daily sales volume was used as the abscissa single product selling price as the ordinate to establish a K-means cluster analysis model. This algorithm was proposed by J.B. MacQueen in 1967. Jiang et al[1] revealed the influence of the evolution of the meso-plastic strain field on the macroscopic response based on the clustered non-uniform transformation field analysis and predicted the mechanical properties of carbon fiber reinforced composites. Zhang et al[2] used K-means cluster analysis to noise the center of the cluster through the planar Laplace mechanism under the premise of satisfying the geographical indistinguishability, and obtain the disturbance position of each location point, to protect the location privacy. Then, the data on the sales volume and category of vegetables per item were extracted and brought into the Pearson correlation analysis model, which was proposed by Carl Pearson in the 80s of the 19th century. Yu et al[3] according to the Pearson theory analysis results, the positive correlation between the moisture content of silty clay and the porosity ratio and plastic limit in this area is moderately correlated, and the positive correlation between the liquid index and the strong correlation degree. Guan Guanglu[4], Pearson correlation analysis was performed on the total runoff control rate, runoff pollution control rate, and road parameters. This is enough to show that our algorithm can be applied to other areas of current research.

In this article, we first replace mosaic leaves, cauliflowers, aquatic rhizomes, nightshades, peppers, and edible fungi with 1, 2, 3, 4, 5, and 6, respectively. Then, the data of vegetable categories and single-day sales volume were standardized by Z-Score, a quadratic parabolic regression model was established to obtain the relationship equation between category and single-day sales volume, and then the single-day sales volume was used as the abscissa single product price as the ordinate to establish a K-means clustering analysis model, and finally, to study the size of the correlation, the data of vegetable single-product sales volume and category were brought into the Pearson correlation analysis model.

2. Data Preparation and Overview

2.1. Data Cleansing

The data studied in this paper are all from China Vegetable Network: http://www.vegnet.com.cn. It is necessary to analyze the distribution law and interrelationship of the sales volume of various categories and single products of vegetables. Firstly, the classification names of vegetables are individually coded, as shown in table 1:

Table 1. Single-heat coding table for vegetable categories

Mosaic	Cauliflower	Aquatic rhizomes	Nightshades	Chili peppers	Edible mushrooms
1	2	3	4	5	6

Replace mosaic leaves, cauliflowers, aquatic rhizomes, nightshades, peppers, and edible fungi with 1, 2, 3, 4, 5, and 6, respectively. The purpose is to make it easier to analyze the sales volume of a single product. Then, the categories of vegetables in the data and the sales volume of single products are matched, and it is found that the single product sales of some products are missing, as shown in the following table2:

Table 2. A summary table of missing values for the number of individual sales

Names	lotus root	Wuhu Green Pepper(2)	Wuhu Green Pepper(2)	Cordyceps flowers (box)
Category	3	5	5	6
The item number	1029000 05116042	1029000 11023648	1029000 11032145	1029000 11011782
Fill in the value	0.4161	0.4074	0.4074	0.3882

For lotus root, Wuhu green pepper (2), Wuhu green pepper (portion), and cordyceps flower (box) The sales volume of the missing items are all replaced by the average sales volume of the single product in their respective categories: 0.4161, 0.4074, 0.4074, 0.3882. Finally, the data after filling in the missing values is visualized, and the line chart of the sales per unit of vegetables in 3, 5, and 6 categories is drawn with MATLAB2022, as shown in the following figure1:

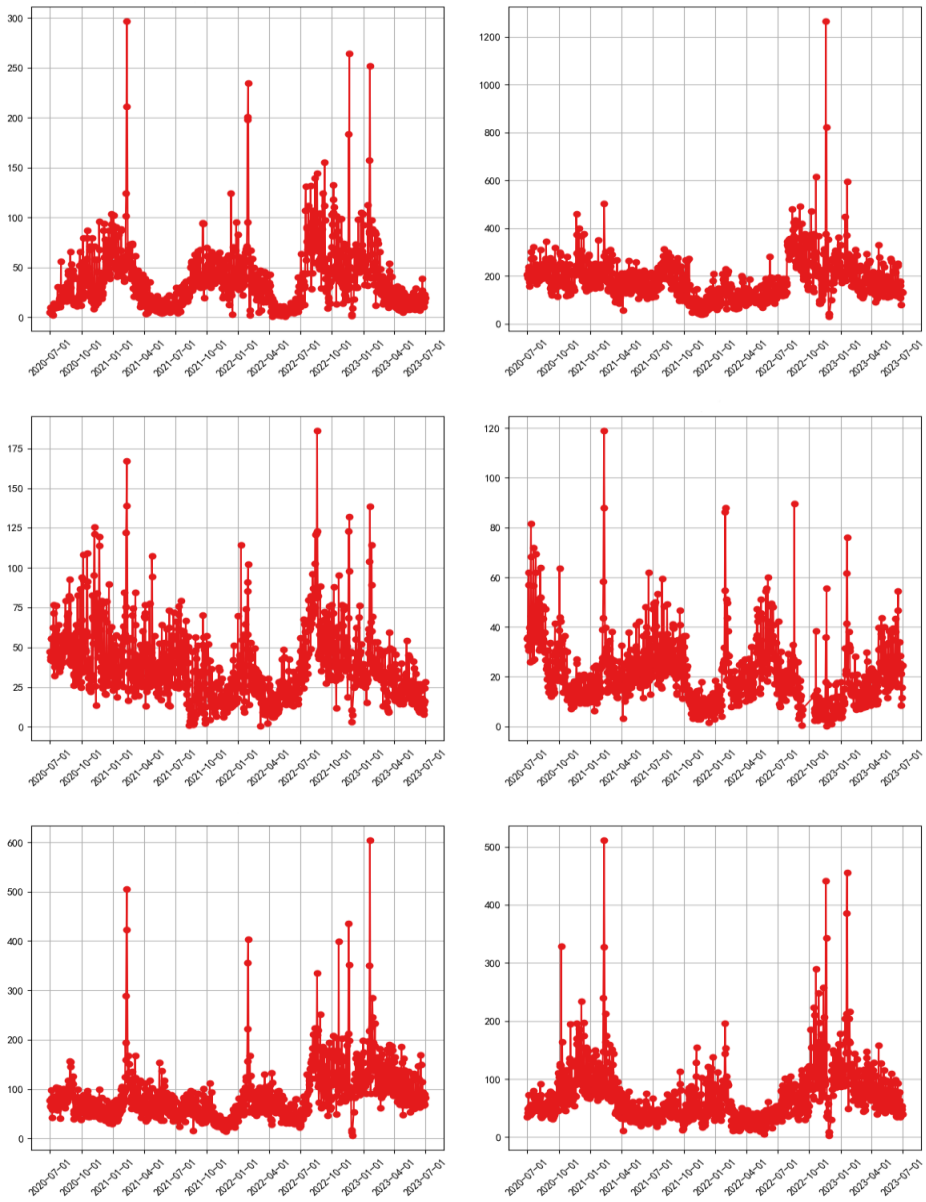


Figure 1. Six different kinds of daily sales can be viewed on the line-line

From the daily sales volume and line charts of aquatic roots, peppers, and foods, it is found that the sales trend of the three categories of vegetables is roughly the same throughout the year, and the sales of the three vegetables are at their highest in the year from September to January of the following year. In the summer of each year, the single-day sales of all three vegetables are at a low point of the year. It shows that people's purchasing power is the same when buying different types of vegetables, and people need to consume a large amount of vegetables in autumn and winter.

2.2. Z-Score normalized processing

Through observation, it was found that the categories of vegetables were divided into 6 categories: 1, 2, 3, 4, 5, and 6, so the sales volume of each type of vegetable was extracted and the average sales volume of each category of vegetables was calculated by Excel. The data were z-score normalized before polynomial regression was performed. The calculated mean, as well as the mathematical formula for Z-Score normalization, is as follows:

Table 3. A table of average sales per item for each category

Categories	1	2	3	4	5	6
Average	0.5980	0.4824	0.6919	0.4996	0.4403	0.5126

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \mu\right)^2}{n}} \tag{1}$$

$$\sigma_j^2 = \frac{1}{m}\sum_{i=0}^{m-1}\left(x_j^{(i)} - \mu_j\right)^2 \tag{2}$$

where σ、μ_j are the standard deviation and the mean, respectively. Therefore, MATLAB calculates the average value of the sales volume of each category after Z-Score standardization, as shown in the following table4:

Table 4. A table of averages for individual items sold by category

Categories	1	2	3	4	5	6
Normal	0.6602	-0.6005	1.6842	-0.4130	-1.0597	-0.2712

3. The process of analysis of the vegetable market

Firstly, a quadratic parabolic regression model was established to obtain the relationship equation between category and single-day sales volume, and then the single-day sales volume was used as the abscissa single product price as the ordinate to establish the K-means clustering analysis model, and finally, to study the size of the correlation, the data of single product sales volume and category of vegetables were brought into the Pearson correlation analysis model.

3.1. Quadratic parabolic regression

In polynomial regression[5], explanatory variables can be quadratic or higher. The parameters of the model are either one-time, so polynomial regression can also be called a linear regression model. The model obtained by regression is a curve, in which the variables are the category of vegetables and the sales volume of vegetables, so the quadratic parabolic regression model is selected, and the mathematical expression is as follows:

$$y = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2 \tag{3}$$

Next, a quadratic parabolic regression is performed on the mean of the sales volume of each category after Z-Score normalization by MATLA2022, and the regression results are shown in the following figure2:

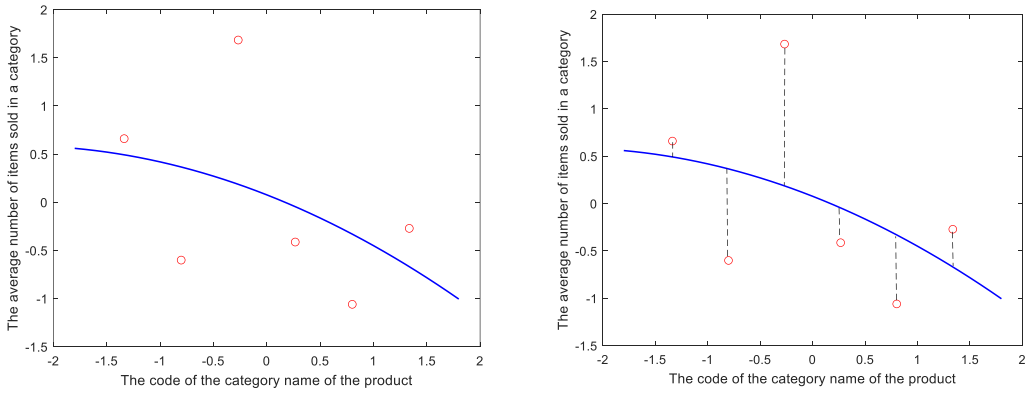


Figure 2. Regression chart of vegetable categories and single item sales

Then, the error analysis values of the quadratic parabolic regression are shown in Table5:

Table 5. Error analysis table for parabolic regression

Original Value	0.598	0.4824	0.6919	0.4996	0.4403	0.5126
Mistake	0.1054	0.1187	0.5123	0.5456	0.769	1.1854
Regression Value	0.4926	0.3637	0.1796	-0.0460	-0.3287	-0.6728

The equation for the quadratic parabola regressed by MATLAB2022 is shown as follows:

$$y = -0.0925 \cdot x^2 - 0.4346 \cdot x + 0.0771$$

Therefore, through quadratic parabolic regression, the relationship between the category of vegetables and the sales volume of vegetables per unit is y=-0.0925x2-0.4346x+0.0771. The absolute error of the regression calculated by Excel is 0.5394, which is caused by the large error of some points, and the overall regression effect has reached the general standard.

3.2. K-means cluster analysis

To explore the distribution law of vegetable categories and single product sales at a deeper level, the daily single product sales volume of vegetables was taken as the abscissa the selling price of vegetable single products was taken as the vertical axis, and K-means cluster analysis was carried out by MATLAB[6]-[8], and the sales characteristics of vegetables from the first half of 2020 to the middle of 2023 were obtained, that is, the characteristics of people buying vegetables. The mathematical formula for K-means clustering is as follows:

$$d\left(X^{(i)}, X^{(j)}\right) = \sum_{k=1}^{p} \sqrt{\left|X_k^{(i)} - X_k^{(j)}\right|^2} \tag{4}$$

Wherein: $X^{(i)}$、$X^{(j)}$ represents the data of group I and group J respectively, and k represents the k-th indicator in each group of data, assuming that there are a total of p indicators. In the beginning, need to determine how many clusters (clusters) there are, iterate through all the data, calculate the distance from each point to the center of n, group the points with the shorter distance, and repeat the cycle until the position of the point does not change. The mean of the data from June 2020 to mid-2023 every 10 days is used as the data for the study, and the results of clustering through MATLA are shown in figure3 below:

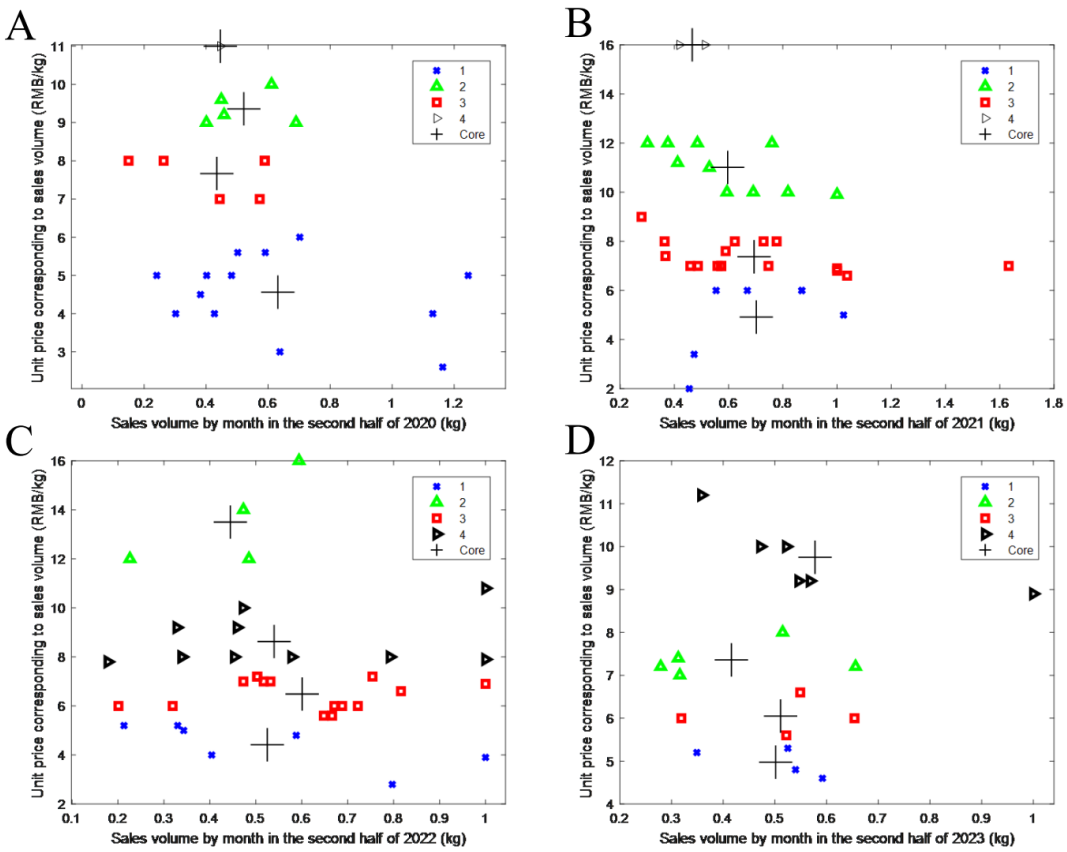


Figure 3. 2020 - 2023Sales in the first half of the year - Unit price cluster chart

The different categories of data in the cluster diagram have been represented in different shapes to make it easier to see the differences between the different categories. To better study the characteristics of data clustering, the clustering clusters of four different periods are counted, as shown in the following table6:

Table 6. K-means cluster summary table

Year	Cluster 1	Cluster 2	Cluster 3	Cluster 4
2020	(0.63,4.56）	(0.52,9.36）	(0.43,7.67）	(0.44,11.0）
2021	(0.70,4.91）	(0.60,11.01）	(0.69,7.37）	(0.47,16）
2022	(0.53, 4.14）	(0.45,13.5）	(0.61,6.49）	(0.54,8.63）
2023	(0.50,4.98）	(0.42,7.36）	(0.51,6.05）	(0.58,9.75）

By analyzing the positions of the four clusters, it is found that the characteristics of the vegetables purchased by people in the four different periods are different, and the characteristics of the vegetables purchased by people in the four periods are drawn by MATLAB, as shown in the following figure4:

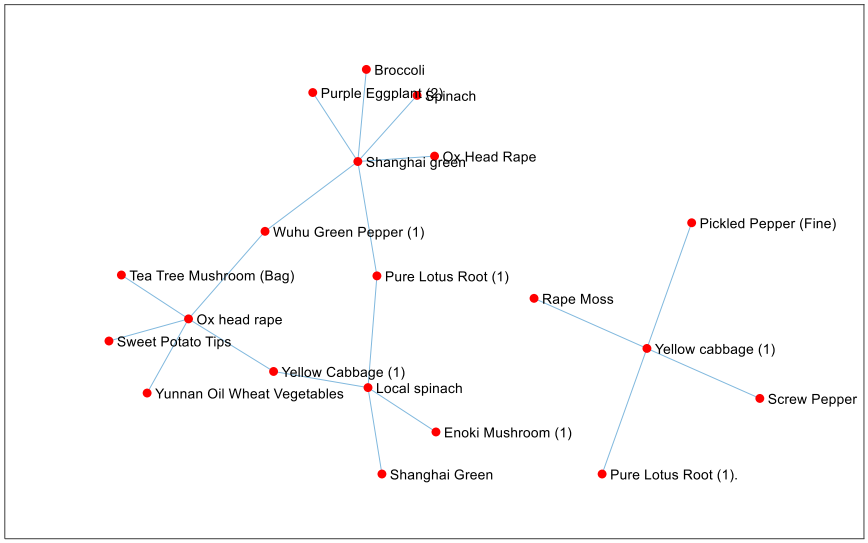


Figure 4. Force chart of people buying vegetables at four different time periods

In the first half of 2020, people tended to buy local spinach, yellow cabbage (1), lotus root (1) and enoki mushroom (1). Throughout 2021, people tend to buy Shanghai greens, spinach, broccoli, pure lotus root (1), purple eggplant (2), and Wuhu green peppers (1). In 2022, people tend to buy Ox Head Oilseed Rape, Yunnan Oily Wheat Rape, Sweet Potato Tips, Wuhu Green Pepper (1), and Tea Tree Mushroom (bag). In 2023, people tend to buy yellow cabbage (1), rape moss, pure lotus root (1), pickled pepper (high-quality), and screw pepper. This phenomenon may be determined by factors such as people's consumption level, natural climate and environment, epidemic situation, plant growth conditions, etc.

3.3. Pearson Correlation Analysis

3.3.1 Determination of normal distribution

Each point of a string of points in the data is a quantile of the data, and the pairing of these points (called the sample quantile points) with the corresponding theoretical quantile to make a scatter plot, if the data follows a normal distribution, then the graph should look like a straight line, otherwise it does not follow a normal distribution. So we used MATLAB2022 to draw a QQ plot of the raw data for Pearson correlation analysis, as shown in 5 below:

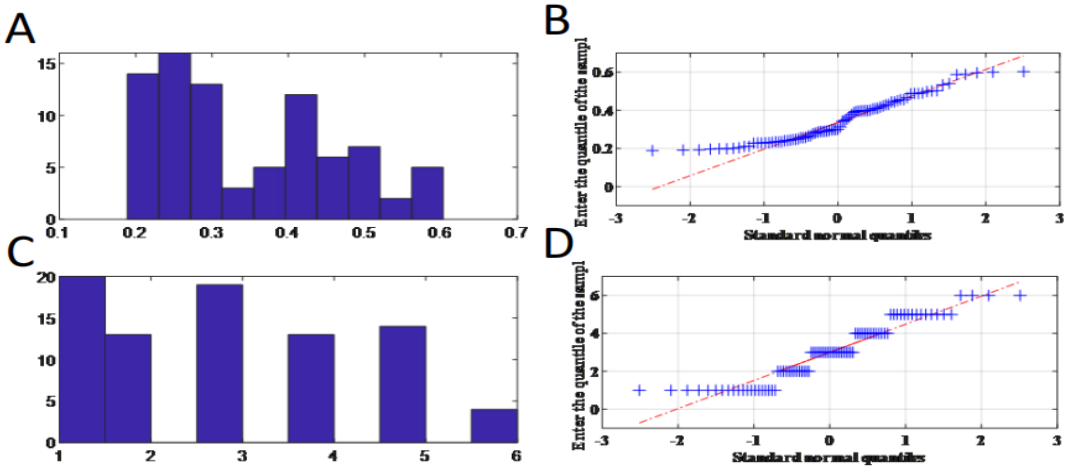


Figure 5. A QQ graph of the data analyzed by sales and vegetable

By observing the QQ chart, the characteristics of the normal distribution of the basic signs of the raw data are discovered, and then the Pearson correlation analysis can be carried out.

3.3.2 Calculate the correlation coefficient

The Pearson correlation coefficient[9]-[10] is essentially a linear correlation coefficient that is used to study the tendency of two sets of data to change and move at the same time. The correlation coefficient > 0 indicates a positive correlation and the correlation coefficient < 0 indicates a negative correlation, the closer the correlation coefficient is to -1 or 1, the stronger their correlation is, and the closer the correlation coefficient is to 0, the stronger their correlation is. The correlation coefficient is the covariance of two variables divided by their standard deviation, and the mathematical formula is as follows:

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X\sigma_Y}=\frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}} \quad (5)$$

where μ is the mean? If the number of samples is represented by N, the above equation can be simplified as:

$$\rho_{X,Y}=\frac{\sum XY-\frac{\sum X\sum Y}{N}}{\sqrt{\left[\sum X^2-\frac{\left(\sum X\right)^2}{N}\right]\left[\sum Y^2-\frac{\left(\sum Y\right)^2}{N}\right]}} \quad (6)$$

Then, the Pearson correlation coefficient and relationship matrix calculated by MATLAB are shown in the following figure6:

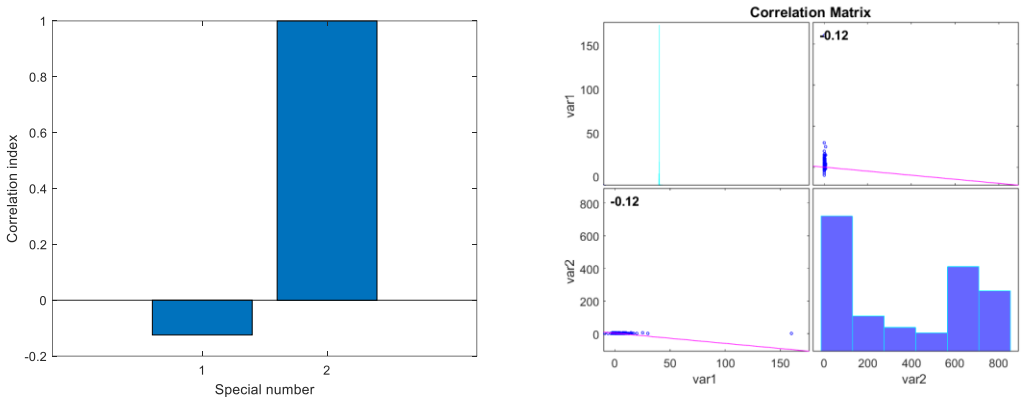


Figure 6. Pearson correlation analysis chart

According to the results of the Pearson correlation coefficient matrix and correlation analysis, the Pearson correlation coefficient of single product sales of vegetable products and vegetable categories is -0.1242, and the Pearson correlation coefficient of vegetable categories and single product sales is 1.0000. This shows that the decision of vegetable category is very dependent on the sales volume of a single product, which shows that the sales volume of different categories of vegetables is very different.

4. Conclusion

Firstly, the categories of vegetables in the data were encoded: mosaic and leaf, cauliflower, aquatic rhizomes, nightshades, peppers, and edible fungi were replaced by 1, 2, 3, 4, 5, and 6, respectively. Then, the data of category and daily sales volume were extracted from the standardized by Z-Score, and a quadratic parabolic regression model was established to obtain the parameters of the relationship equation between category and single-day sales volume as follows: -0.0925, -0.4346 and 0.0771. Then, to further study the distribution of vegetable categories and single-day sales, the K-means clustering analysis model was established by taking the single-day sales volume as the abscissa single product price as the ordinate, and the results showed that people's demand for vegetables was

not fixed in different periods. Finally, to study the size of the correlation, the data of the single product sales volume and category of vegetables were extracted and brought into the Pearson correlation analysis model, and their correlations were -0.1242 and 1.0000. This paper provides a research idea and framework for the field of market analysis, comparing the analysis results of this paper with the actual historical data, and discovering the actual situation of the load, which proves the feasibility of our research method.

References

[1] Jiang Yinan, Chen Juanjuan, Liang Lihua, et al. Clustering non-uniform transformation field analysis of carbon fiber reinforced composites [J]. Chinese Journal of Solid Mechanics, 2023,18(15): 1-4.

[2] Zhang Weiyuan, Tang Wenbing, Chen Yanjie. Research on location privacy method of fusion clustering and differential privacy [J]. Journal of Chifeng University, 2023,5(4): 1-5.

[3] Yu Yi, Liu Sisi, Jiang Jin, et al. Pearson analysis and correlation study of moisture content of silty clay in Yiyang [J].Highway, 2020, 65(7): 7.

[4] Guan Guanglu. Pearson analysis and research on LID technical indicators of municipal roads in Xiamen Jimei Software Park [J].Engineering Construction and Design, 2023, 10(7): 95-99.

[5] Zhang Li, Mi Hong. Mathematics for Middle School Students [J]. Systems Engineering-Theory and Practice, 2010, 30(1): 158-160.

[6] Zhang Hui, Li Kefang, Li Zhizhen. Analysis of College Students' Online Shopping Complaint Behavior Based on K-means Clustering Method [J]. 2021, 78(67): 101-102.

[7] Wang Shihai, An Dong, Liu Bin, et al. A Test Case Screening Method Based on K - Means Clustering and Similarity [J]. Engineering Construction and Design, 2023, 10(7): 95-99.

[8] Luo Xinyi. A comparative study on knowledge diffusion of k-means clustering algorithms——Taking library and information science, management, and economics as an example [J]. Computer Knowledge and Technology, 2023, 19(19): 10-15.

[9] Dutilleul P , Stockwell J D , Frigon D , et al. The Mantel Test versus Pearson's Correlation Analysis: Assessment of the Differences for Biological and Environmental Studies [J]. Journal of Agricultural Biological and Environmental Statistics, 2000, 5(2): 131-150.

[10] Liu W. Gold Price Analysis and Prediction based on Pearson Correlation Analysis [D].Wuhan: Wuhan University, 2023.