Information Systems and Economics (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/infse.2024.050119
ISSN 2523-6407 Vol. 5 Num. 1

Analysis of Merchandise sales by factor, principal component and cluster models

Jiayi Wang, Chunhua Ji

Zhonghua Vocational College of Ynufe, Yunnan University of Finance and Economics, Anning, 650399, China

Keywords: Factor analysis, PCA, K-means clustering

Abstract: The correlation between different categories was determined by analyzing the six categories through factor analysis model, in which the correlation between cauliflower and leafy vegetables was the strongest, and the correlation between cauliflower and eggplant was the weakest. The method used in this paper can correlate the correlation between vegetables, and provide qualitative and quantitative analysis for the sales behavior between different vegetable dealers, and provide reference ideas and suggestions for their reasonable and effective operation. By considering the correlation between goods, their placement can be improved to result in a chain reaction of dish sales and improve overall sales. This can be achieved by placing complementary dishes together, allowing customers to directly purchase items that complement their meal. The goal is to achieve breakthroughs from a single product to the category.

1. Introduction

The article analyzes data from a fresh food supermarket over a three-year period, including purchases, sales records, and purchase prices. The goal is to study the correlation between different vegetable categories and individual products, as well as the distribution patterns and interrelationships within each category. This analysis aims to develop a more effective replenishment plan and pricing strategy for the fresh food supermarket [1].

After researching relevant literature, it is evident that numerous documents discuss the transition from a single product to a category. One example comes from the general manager of Xi'an Craftsmanship Huiteng Enterprise Management Consulting Co. Liang Xianhui, the chief expert of Craftsmanship Marketing Agency, suggests that some industrial enterprises can increase sales of their product line by creating breakthroughs in a single product category. This tactic enhances brand influence and drives sales of other related products. Additionally, retailers can utilize single product breakthroughs to increase sales in relevant categories and further expand the brand's impact. This approach aligns with the strategy of using a single product to drive growth across a category. On the other hand, the retail kiosk can boost sales of associated categories by introducing breakthrough products, which can effectively enhance the brand influence of the chain [2-4]. UFIDA Software Co., Ltd. circulation and retail industry solutions division of the literature also mentioned that the single product management refers to the management of each commodity item as a unit, emphasizing the cost management of each single product, sales performance management; single product management

is the core of the modernization of the management of the chain company's commodities, the commodity group is a combination of a single product; single product management is the basis of the management of the group of commodities, the management of single products to ensure that each type of commodity purchasing, sales and marketing. As well as a listed company in Jiangsu Province, South China's group director assistant on Jia also mentioned from category management to single product management. In the past, the big stores big enough, one-stop service content the more full the better, category roles and structure of the combination is the operation of the skills, category brand and commodity decision-making can be structured to complement each other, and under the small store, small class may be difficult to do all, the system is bound to be upgraded from category management to single product management, management focus and most of the energy to fall back to the single product research [5]. As a result, the correlation between single product and category is particularly important, in the article, we first look for the distribution pattern of each category and single product according to the monthly sales volume, and then, we used the factor analysis model to analyze the relationship between each category. For the single product, due to the number of more than, we based on monthly sales volume of each single product using principal component analysis model to filter out the single product with greater impact, and then using K-means clustering, the single product was classified. Finally, the correlations between different categories and different individual items and the interrelationships between different similar categories and individual items were derived. The significance of this work is that this fresh food supermarket can reduce the loss caused by too much stock or because of too high pricing according to the conclusion drawn, and better system of replenishment and pricing strategy [6], so as to maximize the benefits, and at the same time, it can provide a reference for other fresh food supermarkets.

2. Relevant model

2.1 Factor analysis model structure

FFactor analysis (FC), also known as factor analysis, is a data analysis technique based on correlation, which is a dimensionality reduction method based on a large number of observations [7]. Its main purpose is to explore a certain structure hidden behind a large amount of observed data, to find the common factors of a group of variable changes, and to categorize variables with the same essence into one factor, which can reduce the number of variables, and can also test the hypothesis of the relationship between variables [8].

Based on the sales volume of different categories, and factor analysis formula (1), formula (2) we calculated the cumulative contribution of the three factors, and got the contribution of the sales volume of different categories. The specific steps are as follows:

There are three possible correlated variables $X_1, X_2, X_3$with three independent common factors $F_1, F_2, F_3$and one special factor e_i(i=1,2,3)， $\varepsilon_1, \varepsilon_2, \varepsilon_3$are uncorrelated with each other and with F_j(j=1,2,3)are uncorrelated with each other, each of theX_i(i=1,2,…,p) can be represented linearly by the common factor and its own corresponding special factor.[5] Then the mathematical model of factor analysis is:

$$\begin{aligned} X_1 &= a_{11}F_1 + a_{12}F_2 + a_{13}F_3 + \varepsilon_1 \\ X_2 &= a_{21}F_1 + a_{22}F_2 + a_{23}F_3 + \varepsilon_2 \\ X_3 &= a_{31}F_1 + a_{32}F_2 + a_{33}F_3 + \varepsilon_3 \end{aligned} \tag{1}$$

So the loading factor is:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{bmatrix} \tag{2}$$

2.2 Principal Component Analysis Model

PCA (Principal Component Analysis), or Principal Component Analysis method, is one of the most widely used algorithms for dimensionality reduction of data [9].The main idea of PCA is to map n-dimensional features onto k dimensions, which are brand new orthogonal features also known as principal components, which are k-dimensional features reconstructed on the basis of the original n-dimensional features [10-11]. According to the sales of individual products, and the main hierarchical analysis formula (3) filtered out the sales of varieties with a large cumulative contribution in different categories. PCA specific calculation step process is as follows:

First derive the coefficient matrix R:

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pm} \end{bmatrix} \tag{3}$$

Secondly, the cumulative contribution of the first l principal components, S, was computed.

$$S = \frac{\sum_{k=1}^{l} \lambda_k}{\sum_{i=1}^{p} \lambda_i} (l = 1,2,\cdots,p)$$

2.3 K-means cluster analysis model structure

The k-means clustering algorithm (k-means clustering algorithm) is an iterative solution of cluster analysis algorithm, the steps are, pre-dividing the data into K groups, then randomly select K objects as the initial clustering centers, and then calculate the distance between each object and the various seed clustering centers, and assign each object to the closest clustering center to it [12-14]. The cluster centers and the objects assigned to them represent a cluster [15].

Based on the selected data and the Min's distance formula (4) in the Euclidean distance using Matlab to construct the cluster analysis model of different individual products respectively, finally we get the combination of different individual products. The specific analysis is as follows:

$$d_{ij} = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^q\right)^{\frac{1}{q}} \tag{4}$$

When q = 13

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots \ldots + (x_{ip} - x_{jp})^2} = \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2\right)^{\frac{1}{2}}$$

3. Results

3.1 Analysis of the results of the factor analysis model

The correlation coefficients calculated from the factor analysis, as shown in Table 1 correlation coefficient matrix, can be concluded that the strongest correlation among the six categories is between cauliflower and cauliflower leaf, the weakest correlation is between cauliflower and eggplant, and eggplant is weakly negatively correlated with cauliflower leaf, chili peppers, edible mushrooms, and cauliflower, and weakly positive correlation is found among the other categories.

Table 1: Correlation coefficient matrix

1	0.7469	0.4841	-0.0345	0.6238	0.5461
0.7469	1	0.4721	0.0581	0.4255	0.4287
0.4841	0.4721	1	-0.4660	0.4589	0.6258
-0.0345	0.0581	-0.4660	1	-0.1914	-0.4089
0.6238	0.4255	0.4589	-0.1914	1	0.5775
0.5461	0.4287	0.6258	-0.4089	0.5775	1

The results of the three-factor analysis were obtained in Table 2 below, and the cumulative contributions of the three factors were calculated as 25.9956, 48.9400, 71.0780.

Table 2: Results of factor analysis

factor1	factor2	factor3
0.6319	0.0898	0.6124
0.9177	0.0437	0.2664
0.4018	0.6904	0.2741
0.1188	-0.7107	-0.0747
0.2381	0.2314	0.7391
0.2932	0.5712	0.5054

3.2 Results of principal component analysis and cluster analysis

Through the principal component analysis, the single products that have a greater cumulative contribution to sales in different categories were screened out, as shown in Table 3.

Table 3: Principal component analysis

philodendron	cauliflower	aquatic root type	eggplant	capsicum	edible mushroom
15.7617	31.90767	22.19136	27.19321	29.97659	17.56998
25.40232	58.17361	38.37556	47.1882	40.24598	27.99799
34.47291	76.35581	48.72035	61.36015	50.00908	36.00705
41.3195		57.29609	72.56695	57.92049	43.5424
47.81431		64.20312	82.46131	64.68069	50.27106
53.79902		70.46946		70.67319	56.17185
59.0467		76.35914		75.39195	60.82738
63.53146				79.07307	65.34742
67.56307					68.91083
71.23009					72.21091
74.51272					75.21933
77.65554					78.0138
79.96956					80.24541

On the basis of this principal component analysis, using K-means cluster analysis, the single product is sold better when it is divided into the following four categories, and the specific results are shown in Figure 1.

The first category: Chrysanthemum coronarium, Caidian quinoa;

The second category: Shanghai green, radish leaves, Niushou rape;

The third category: Niushou lettuce, Chinese cabbage, kale leaves, choy sum, mullein, pea tips, Wuhu green peppers(1), flowering eggplants, Chinese cabbage, Sichuan red parsnips, and purple bell peppers;

Fourth category: green stalks and scattered flowers, Honghu lotus root (pink lotus root), net lotus root (2), fresh lotus root ribbons (bag), water chestnuts (portion), high melon, Honghu lotus root (crisp lotus root), net lotus root (3), purple eggplants (2), green eggplants (1), purple round eggplants, big long eggplants, red sharp peppers, millet peppers, screw peppers, red string peppers, red bell peppers (1), water peppers (orange), Wuhu green peppers (1), combination of peppers series, Xixia flower mushrooms (1), apricot mushroom(1), shimeji mushroom(bag), enoki mushroom(bag)(1), silver fungus(dos), monkey head mushroom, fresh fungus(1), chanterelle mushroom, black porcini mushroom, seafood mushroom(bag)(1), shiitake mushroom, xixia shiitake mushroom(1), tea tree mushroom(bag);

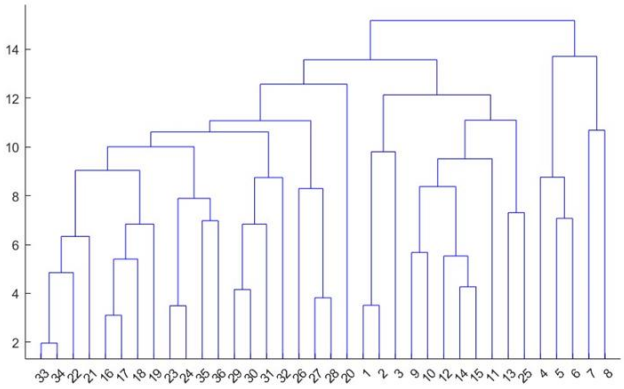

Figure 1: Clustering lineage diagram

4. Conclusions

For the category, we use factor analysis model to analyze the monthly sales volume of each category in this fresh food supermarket, and conclude that the strongest correlation among the six categories is cauliflower and foliage, the weakest correlation is cauliflower and eggplant, and eggplant is weakly negatively correlated with foliage, chili peppers, edible mushrooms, cauliflower, and weakly positively correlated with the other categories. For the single product, we used K-means clustering after screening the single product with high influence through principal component analysis, and came to the following conclusions:

The first category: Chrysanthemum coronarium, Caidian quinoa;

The second category: Shanghai green, radish leaf, and Niushou oilseed rape;

The third category: Niushou lettuce, chard, kale leaves, choy sum, wood ear greens, pea tips, Wuhu green peppers (1), flowering eggplants, Chinese cabbage, Sichuan red parsnips, and purple-tipped peppers;Fourth category: green stalks and scattered flowers, Honghu lotus root (pink lotus root), net lotus root (2), fresh lotus root ribbons (bag), water chestnuts (portion), high melon, Honghu lotus root (crisp lotus root), net lotus root (3), purple eggplants (2), green eggplants (1), purple round eggplants, big long eggplants, red sharp peppers, millet peppers, screw peppers, red string peppers, red bell

peppers (1), water peppers (orange), Wuhu green peppers (1), combination of peppers series, Xixia flower mushrooms (1), apricot mushrooms(1), shimeji mushrooms(bag), enoki mushrooms(bag)(1), silver fungus(dos), monkey head mushrooms, fresh fungus(1), chanterelle mushrooms, black porcini mushrooms, seafood mushrooms(bag)(1), shiitake mushrooms, Xixia shiitake mushrooms(1), tea tree mushrooms(bag); From this categorization we can conclude that the best sales are achieved when dishes are sold in the above categorized combinations.

Distribution patterns and interrelationships of sales volume of each category and each individual product. For the category: we constructed a line graph of the monthly sales volume of each category, through the line graph can visualize the monthly sales volume distribution pattern of each category, the six categories of vegetables in the leafy flowers, peppers, edible fungi, aquatic roots, cauliflower, the five categories of vegetables have a similar monthly sales volume line graph, and the five categories 2020-2022 three years of the monthly sales volume peaks have appeared in the vegetable supply varieties rich April through October. The trend pattern of the line graph shows that the five categories of foliage, peppers, edible mushrooms, aquatic roots, and cauliflower have a strong positive correlation. Therefore, the sales volume of eggplant is negatively correlated with the other five categories.

For a single product: for a single product between the same categories, this paper use principal component analysis to construct a line graph based on the structure of the monthly sales volume after the use of principal component analysis handsome, as shown in Fig. 2.

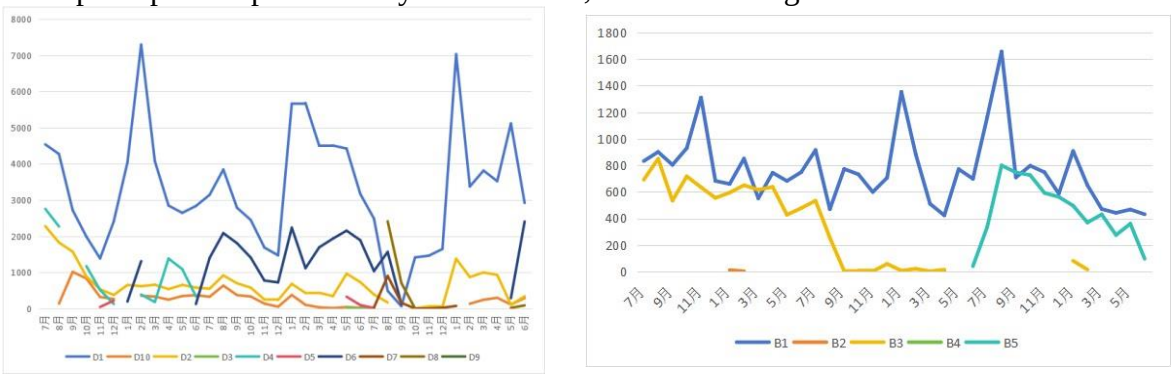

Figure 2: Two single product sales charts

Where the left graph represents the graph of sales volume over time for 10 individual items in the eggplant category and the right graph represents the graph of sales volume over time for 5 individual items in the cauliflower category. As the division pattern of the above graph, in the case of the cauliflower category, this paper can conclude that the sales volume of item B1 has a better trend throughout the year in relation to the demand of customers, and due to the impact of the epidemic, the sales volume of item B2 plummeted after May 2022, while the demand of item B5, from July 2022 to May 2023, increased again.

Outlook of the results. For this problem, when this paper analyze the relationship between categories, we summarize them into distribution line graphs based on monthly sales, which can more intuitively see the general relationship between categories, and then use factor analysis to derive the specific strength of association between categories, so that the results of the model analysis have sufficient accuracy. When analyzing the relationship between individual products, this paper first establish the method of principal component analysis to find the contribution rate and cumulative contribution rate of individual products in each category, and then further establish the cluster analysis model, so as to arrive at the accurate correlation relationship between individual products, and simplify the complexity of the relationship between individual products. It is hoped that in the future, the fresh food supermarket will be able to make better purchase and pricing strategies based on this conclusion, so that the fresh food supermarket can achieve the maximum benefit, avoiding losses caused by factors

such as excess products or overpricing and low sales, and hopefully providing references to other practitioners in the industry.

References

[1] LIN Liqian. Research on Cost Control of Fresh Food Retail Enterprises--Taking Supermarkets as an Example[J]. Mass Investment Guide, 2023(14):185-187.

[2] Wang Zhijin. Using single product as a breakthrough to pry category growth[J]. China Drugstore, 2022(10):68-69.

[3] Wang Yangxue, Jiao Jinyuan. Practical research on non-oil business operation system based on single product management [J]. Chemical Management, 2020(31):36-37.

[4] Commentator. Innovative product "mutual sales" model[J]. Pivot Point, 2023(10):1.

[5] Cang Chun. Parallel single product, the core play of sales improvement[J]. China Drugstore, 2022(09):32-33.

[6] An Kui. Research on cooperative optimization of M fresh supermarket supply chain inventory [D]. Guizhou University, 2022.

[7] J. K N, N. Z S, Jeffrey D N, et al. Detection of differential depressive symptom patterns in a cohort of perinatal women:an exploratory factor analysis using a robust statistics approach[J]. e Clinical Medicine, 2023, 57.

[8] Zhang MX, Yang Y. Research on financial risk evaluation and control of listed enterprises in Heilongjiang province based on factor analysis [J]. Modern Auditing and Accounting, 2023(09):25-27.

[9] YU Jing, JIANG Anlin, LIU Liang et al. Research on aerodynamic shape parameterization method based on PCA dimensionality reduction [J/OL]. Journal of Aeronautics: 1-17

[10] Chao M A , Sen C , Xiao-Bo L ,et al.Evaluation on simulative transportation and shelf quality of blueberries by different treatments based on principal component analysis[J].Food Science and Technology, 2018.

[11] JI Li, LIU Xiaoran, WU Qiang et al. Establishment of a prediction model for the first flowering period of waxberry flowers based on PCA [J]. Journal of Southwest Normal University (Natural Science Edition), 2022, 47(10):59-66.

[12] Kang Glimmer. Research on subclassification of glass artifacts based on improved GDBT and K-means clustering [J]. Information Technology and Informatization, 2023(07):149-152.

[13] SUN Lin, LIU Menghan. K-means clustering based on adaptive cuckoo optimized feature selection[J/OL]. Computer Applications: 1-13

[14] Wu Huihui, Yuan Zhe, Hui Xiaojian et al. Research analysis of Olympic awards based on K-means clustering[J]. Modern Information Technology, 2023, 7(15):136-140.

[15] Yinbo Xu, Yang Yu. Detection of abnormal data in ship communication network based on K-means clustering[J]. Ship Science and Technology, 2023, 45(16):169-172