

Multi-modal Analysis of Water Milk Consumption Preferences and Influencing Factors——Based on Shanghai's Generation Z Crowd

Jianrong Li *

Department of Economic statistics, Shanghai Polytechnic University, Shanghai China, 201209

* Corresponding author: 18077265201@163.com

Abstract. This study aims to analyze the consumption preference and influencing factors of Buffalo Milk among Shanghai's Generation Z population, to provide marketing strategy references for dairy companies, and to promote the healthy development of Shanghai's Buffalo Milk market. In this paper, we first take the Generation Z population as the main body through network survey and literature survey, and select potential consumers who will buy buffalo milk from them, and then recover the questionnaires to conduct follow-up surveys, organize and analyze the existing buffalo milk industry data, and derive the overall development of Shanghai's Generation Z buffalo milk market, and compare the customer's loyalty to the brand by using the retention index in order to explore the brand equity of each brand. The brand equity of each brand is explored by utilizing the retention index to compare customer loyalty. Secondly, we collect basic data on buffalo milk consumption of Generation Z through online questionnaire survey based on K-Means clustering algorithm for feature extraction, and use word cloud to draw user profiles to find out the consumption preference of buffalo milk crowd. Then, the logistic model is used to explore the significant influencing factors that affect consumers' purchase of buffalo milk. Finally, we provide reasonable and effective suggestions for the "broken circle" problem of buffalo milk.

Keywords: Buffalo Milk, Consumers, Logistic Regression, K-means Cluster Analysis, Customer Profiling.

1. Introduction

1.1. Background

In the milk-based dairy market, the emergence of niche milk has brought new choices, with the advantages of taste, milk source, and so on, more and more appear in front of consumers. However, the current "niche milk" market is in the wild growth stage of the dry good and bad, there are a lot of problems and chaos [1]: First, insufficient production capacity; second is the production management level to be improved; third is suspected of excessive publicity, do not focus on research and development, focus on marketing. All of the above may become constraints to the development of niche milk [2]. Based on the preliminary pre-survey, this project takes "Generation Z" consumers as the research object, focuses on exploring the current situation of the buffalo milk market [3], consumer behavioral characteristics, etc., so as to put forward suitable sales strategies for buffalo milk to "break the circle". The research objectives are as follows:

- 1) To understand the current status of the Shanghai buffalo milk market at the merchant level and analyze the brand equity (brand loyalty, brand awareness, etc.) of each brand in the Shanghai buffalo milk market;
- 2) Exploring the behavioral characteristics of "Generation Z" consumers and their influencing factors at the consumer level.
- 3) Providing sound and effective advice on "breaking the loop" on buffalo milk.

2. Materials and methodologies:

2.1. Data collection

This survey used the network questionnaire method to collect questionnaire information, the questionnaire consists of 23 questions [4], the survey was issued 310 questionnaires, a total of 306 questionnaires were recovered, excluding the data samples used to fill in the time less than 60 seconds, the actual effective questionnaire is 300, the validity rate is 98%. Then SPSS statistical software was used to test the validity of the valid 300 questionnaires using the KMO (Kaiser-Meyer-Olkin) test statistic and Bartlett's spherical test. In SPSS statistical software analysis - dimensionality reduction - factor analysis - description - correlation matrix - KMO test & Bartlett's spherical test, the variables were put into the factor analysis to do the KMO and Bartlett's spherical test, and the results were obtained as Table 1:

Table 1. Questionnaire data test table

KMO Sampling Suitability Measure	0.681
Bartlett's test of sphericity approximates the chi-square	4664.47
Bartlett Spherical Degree of Freedom	957
Bartlett's spherical significance	0.00

The KMO statistic is 0.681 close to 0.7, the correlation between the variables is relatively strong, the partial correlation is weak, and the factor analysis is not bad. The value of the statistic in the Bartlett's test of sphericity is 4664.487, and the P-value is approximately equal to 0. Therefore, the original hypothesis is rejected, and it is considered that the correlation coefficient is not likely to be a unit array, i.e., there is a correlation between the original variables, and it is suitable for the factor analysis and the structure of the questionnaire is well-designed. In conclusion, the questionnaire and sample used in this survey are reasonable and reliable and can be used for further analysis.

2.2. Data basics

Among the consumers in the research, 36% of the men and 64% of the women have purchased buffalo milk, their ages are mostly concentrated in 18-25 years old, their occupations are mostly employees of enterprises and institutions and students, the amount of money they spend on buffalo milk per month is mostly concentrated in 50-75 or more than 75 RMB, and their education levels are mostly concentrated in universities. The basic attributes of users are shown in the table.

Table 2. User basic information table

	Project	Sample size (statistics)	Percentage
Sex	Males	108	36%
	Females	192	64%
Age	Under 18	80	26%
	For ages 18-24	128	42%
	For ages 25-28	92	32%
Careers	students	112	38%
	migrant workers	34	11%
	Enterprises and institutions	82	28%
	Management	31	11%
	self-employed	23	7%
	workless	18	5%
Level of education	High school and below	57	19%
	University colleges and universities	87	29%
	Graduate students and above	156	52%

2.3. Methods

2.3.1. Research ideas

In order to achieve the above-mentioned business objectives, this project takes "Generation Z" consumers in Shanghai as the research object, designs a questionnaire to explore the characteristics of consumer behavior, and provides effective suggestions for "breaking the circle" of buffalo milk. The specific content is as follows:

- 1) Conducted research on the market of buffalo milk stores in Shanghai to understand the sales status of the buffalo milk brand;
- 2) Using the retention index, the customer loyalty of each buffalo milk brand was compared to obtain the brand with the largest retention index, and interviews were conducted with customers to gain a deeper understanding of the reasons why the brand was able to retain customers.;
- 3) Questionnaire method is used to obtain user data, K-Means clustering method is used for feature extraction, and word cloud method is used to draw user profiles;
- 4) Based on the logistic regression in the analyzed questionnaire to explore the influencing factors of buffalo milk purchase frequency finally based on the above conclusion buffalo milk merchants proposed the optimal marketing strategy;

2.3.2. Introduction to Methods

1) K-means Clustering: K-means clustering algorithm was proposed by J.B. MacQueen in 1967, it is the most widely used algorithm among the division-based clustering algorithms, and it is also one of the distance-based clustering algorithms. Distance-based clustering algorithms mainly utilize the distance metric as an evaluation criterion [5]. K-means clustering algorithm is the most typical algorithm among the above algorithms, and it is also the most commonly used clustering algorithm. The basic idea of the K-means algorithm is as follows: firstly, the user specifies an appropriate number of clusters, K; secondly, the algorithm automatically selects K nodes arbitrarily in the dataset to be the center of clustering, and forms K clusters; again, 40 iteratively assign nodes to each clustering center and recalculate the clustering center; finally, repeat the above process until the clustering center no longer changes to return the best clustering results. During each iteration, nodes should be assigned to the nearest cluster centers as much as possible, so as to minimize the value of the criterion function and eventually converge to a very small value that no longer changes.

2) Logistic regression model: logistic regression is a generalized linear regression model for the analysis of discrete random variables [6] in the general form:

$$\ln[p_i / ((1 - p_i))] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \cdots + \beta_m X_{mi} \quad (1)$$

Where Y_i is the binary response variable, taking the value of 0 or 1, X_{1i} , X_{2i} , ..., X_{mi} is the response variable Y have an effect on the m independent variables, in the m independent variables under the action of the probability of "success" for p_i that is:

$$p_i = P\{Y_i = 1 | X_{1i}, X_{2i}, \cdots, X_{mi}\} \quad (2)$$

The probability of "failure" under the action of m independent variables is $1 - p$, i.e.:

$$(1 - p_i) = P\{Y_i = 0 | X_{1i}, X_{2i}, \cdots, X_{mi}\} \quad (3)$$

3. Modeling & Analysis

3.1. Analysis of basic information of the questionnaire

According to the survey data, nearly half of the consumers learned about buffalo milk through popular APPs such as Jitterbug and Xiaohongshu, while a few consumers learned about buffalo milk through elevator buildings, subway buses and poster campaigns.

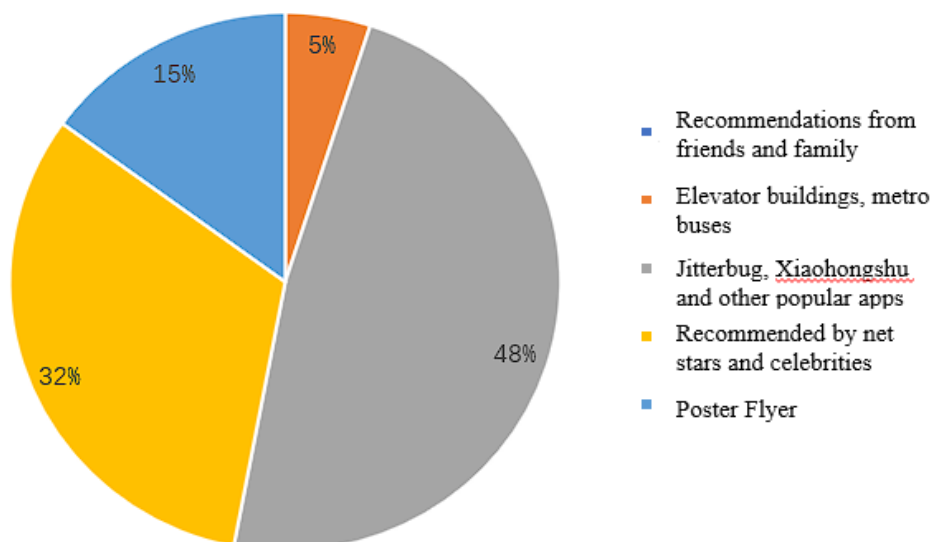


Figure 1. Consumer Perception of Buffalo Milk Channels

After consumer motivation, the majority of consumers preferred to purchase buffalo milk through platforms such as Meituan Preferred, followed by about 28% of consumers who purchased buffalo milk through platforms such as Jitterbug and Shutterbug, and a small number of consumers who purchased buffalo milk through platforms such as Taobao, Jingdong, and store supermarkets, which amounted to 19% and 10%, respectively.

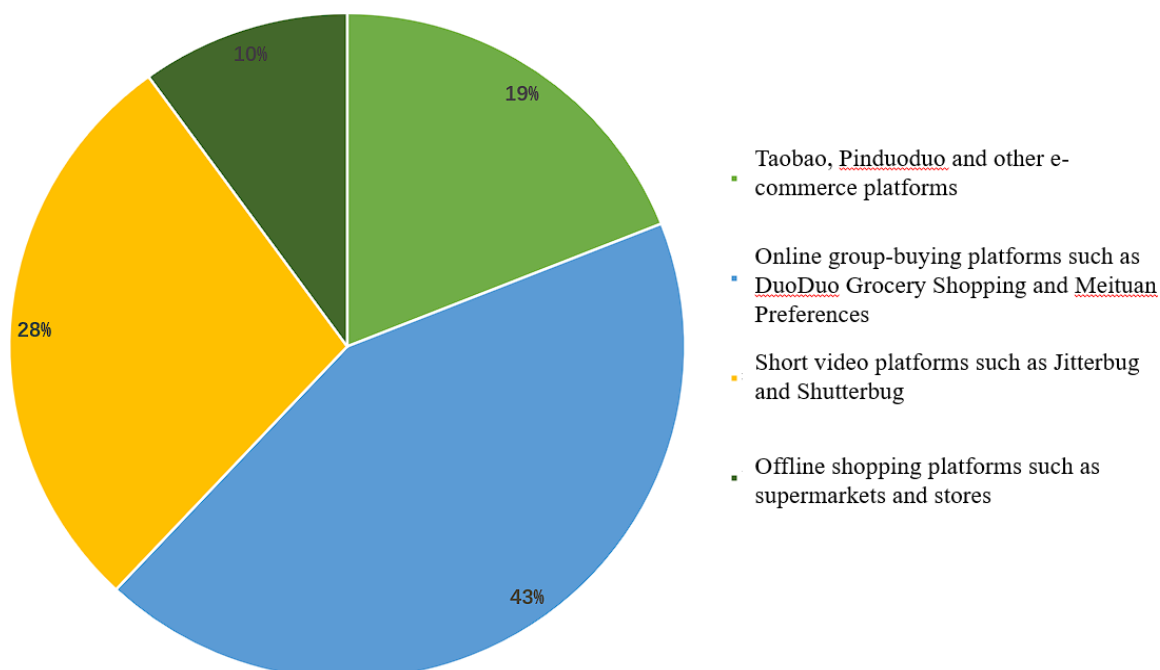


Figure 2. Consumer channels for purchasing buffalo milk

It can be seen that Generation Z consumers are more closely in touch with the Internet, and most of them learn about buffalo milk through online platforms and purchase buffalo milk through online channels.

In terms of the distribution of the amount of money consumers spend on buffalo milk each month, in general, each group spends more than 75 yuan or more on buffalo milk each month, and the percentage share is close to 50%. Graduate students, due to their higher education and income levels, pay more attention to better quality milk, and tend to spend more than \$75 per month on buffalo milk. Undergraduate students, whose income is less compared to graduate students, tend to spend more on \$50-75. High school students and those below high school have lower education levels and lower incomes, and their purchases of buffalo milk tend to be in the range of \$25-\$50.

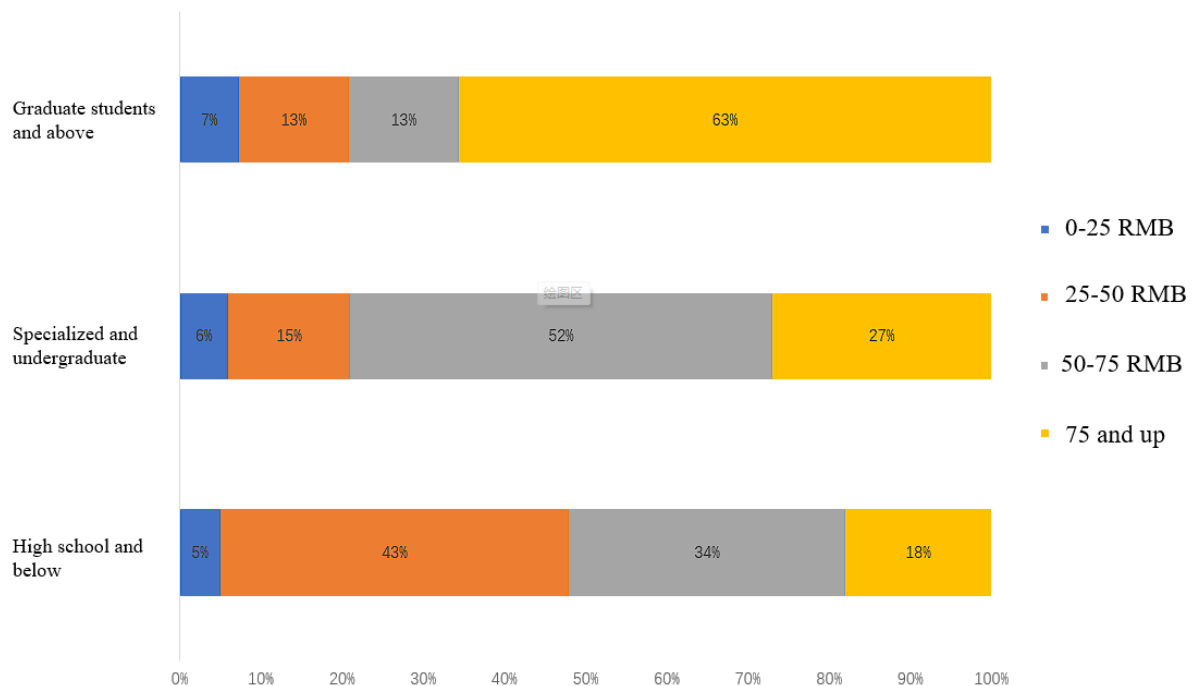


Figure 3. Educational qualifications and purchase amounts

3.2. Consumer Group Analysis

In this research, the frequency of drinking and purchase amount were analyzed by ANOVA test through SPSS. The results of the test are shown in Tables 1-2, and both variables are significantly different, indicating that the process of cluster analysis effectively focuses on users with similar characteristics, and that the characteristics of interaction behaviors vary greatly between different user types.

Table 3. ANOVA results

ANOVA						
	clustering		inaccuracies		F	significance
	mean square	degrees of freedom	mean square	degrees of freedom		
Drinking frequency	16.584	3	.249	58	66.604	.000
Purchase amount	14.950	3	.233	58	64.279	.000

The final clustering clusters the users into four categories, which are divided into important value users (higher drinking frequency and higher single purchase amount), common retention users (higher drinking frequency but lower single purchase amount), common development users (lower drinking frequency but higher single purchase amount), and common maintenance users (lower drinking frequency and lower single purchase amount) [7]. The information characteristics of the four clustered groups are shown in the table below:

Table 4. Final Cluster Centers

cluster	Cognitive situation	Purchase Characteristics
Important Value User	They were introduced to buffalo milk through recommendations from friends and family and popular apps, and love the novel taste of its ice cream flavors.	Purchase channels are mainly stores, supermarkets, etc.; the frequency of purchasing buffalo milk is very high, and the amount of a single purchase is high; the three aspects of its brand awareness, product quality and packaging are valued.
General Retention User	They are recommended by popular APPs as well as Netflix and celebrities, and Netflix celebrity endorsements are more likely to increase the purchase desire of this category of consumers who like the novel flavor of their ice cream.	Their main purchasing channels are online group-buying platforms; higher purchasing frequency; higher single purchase amount when purchasing buffalo milk is more important to brand awareness, milk freshness, price and other aspects; like to save environmentally friendly packaging.
General Development User	The most exposure to buffalo milk is through popular APP platforms such as Jitterbug Xiaohongshu and poster leaflets	The main purchase channels for Taobao, Pinduoduo, Jingdong, as well as jittery, fast hand and other platforms, the purchase of buffalo milk frequency is low, but the single purchase amount is higher concentrated in the 25-50 yuan, the main purchase of high-calcium milk, most of the importance of the buffalo milk brand and product quality, like the cute and weird type of packaging.
General Maintenance User	The most exposure to buffalo milk was through Netflix and celebrity publicity, with a fondness for buffalo's ice cream flavors.	The main purchasing channels for jittery, fast hand and other platforms, the purchase frequency is low, the purchase amount is not high, most of the emphasis on the sense of experience of the purchase process of buffalo milk, packaging, do not care about the brand, generally buy some not famous brands of buffalo milk packaging style do not care.

3.3. Analysis of Factors Influencing the Frequency of Purchase of Buffalo Milk

3.3.1. Model Variable Settings

The corresponding vector Y was set as the frequency of buffalo milk purchase, and the binary categorical response variable of buffalo milk purchase frequency was set, with the options "occasional purchase", "never purchased" and "rarely purchased" defined as low, and the option "often purchased" defined as high [8]. "Defined as low frequency of purchasing buffalo milk, and the option "often purchased" defined as high frequency of purchasing buffalo milk [8].

Gender, monthly income, time of liking buffalo milk, type of buffalo milk drink purchased, protein content of buffalo milk, price, and the degree of influence on sleep were modeled as independent variables X. In this case, since gender, time of liking buffalo milk, and type of buffalo milk drink were unordered subtyped variables, we defined them to be dealt with as dummy variables [9].

Table 5. Explanatory Table for Logistic Model Variables

Variable name	Variable symbol	Variable definitions
Frequency of buffalo milk purchases	Y	0: Low frequency 1: High frequency
Sex	X1	0: Females; 1: Males
Drinking time	X2	1: Morning 2: Lunch; 3: Dinner; 4: Any time
Type of buffalo milk	X3	1: high-calcium 2: pure milk. 3: No apparent tendency.
Protein contents	X4	1: 2.5-2.8g. 2: 2.8-3.5g. 3: 3.5-3.8g.
Price	X5	1: 0-25RMB. 2: 25-50RMB. 3: 50-70RMB. 4: 75RMB and above.
Monthly salary	X6	1: Less than1500RMB. 2: 1500-3000RMB. 3: 3000-8000RMB. 4: 8000RMB and above.
Sleep effects	X7	1: No effect at all. 2: A little bit. 3: Whenever there's an impact. 4: Fast asleep.

Summarizing yields the following expression:

$$\begin{aligned}
 &Ln\{pi/(1 - pi)\} \\
 &= \beta_0 + \beta_1 * Sex + \beta_2 * Drinking\ time + \beta_3 * Type\ of\ buffalo\ milk \\
 &+ \beta_4 * Protein\ contents + \beta_5 * Price + \beta_6 * Monthly\ salary + \beta_7 \\
 &* Sleep\ effects
 \end{aligned}
 \tag{4}$$

3.3.2. data analysis

The model was terminated at the 8th iteration before the maximum number of iterations for parameter estimation and the model fit was considered successful. The model summary table shows that the Nagelkerke R-squared value is between 0 and 1, here 0.689, therefore, the model is well fitted.

Table 6. Summary of models

Steps	-2 log-likelihood ratio	Calcos Snell R-square	Nagelkerke R-squared
1	26.668a	0.546	.689

The original hypothesis of the Hosmer-Lemeshow test is that the distribution of observed frequencies is not significantly different from the distribution of expected frequencies. At this point $p=0.857>\text{significance level } \alpha=0.05$, so the original hypothesis is accepted, i.e., the discriminant model is considered to be well-fitted.

Table 7. Hosmer-Lemeshaw test

Steps	Chi-square	Degrees of freedom	Significance
1	6.589	8	0.857

In summary, although the logistic regression function can be used to build a discriminant model, there are non-significant discriminant variables in the regression function, so the logistic regression model needs to be improved.

Table 8. Logistic model results

Variable names	Coefficient of regression	Standard error	P-value
X2 Drinking time	.053	0.0728	0.0122
X4 Protein contents	.031	0.1455	0.0130
X6 Monthly salary	.201	0.1631	0.0245
X7 Sleep effects	-.012	0.1666	0.0181
β_0	-.187	0.6240	0.0050

The final logistic regression model is:

$$\ln \left[\frac{p}{1-p} \right] = 0.053 * \text{Drinking time} + 0.031 * \text{Protein contents} + 0.201 \text{Monthly salary} - 0.012 \text{Sleep effects} - 0.187 \quad (5)$$

From the results, it can be seen that the p-value of buffalo milk protein content, time, monthly income and level of sleep is less than 0.05, which means that the above variables are considered to have a significant effect on the frequency of purchase of buffalo milk by consumers. From the regression coefficients, it can be seen that Sleep affects the frequency of consumer purchases to decrease, in contrast to monthly income, buffalo milk protein content, and time, which causes positive feedback on the frequency of consumer purchases [10].

3.3.3. Description of analytical results

1) In summary, we conclude that the effect of consumer age, gender, and drink type on the frequency of buffalo milk purchases is not significant. It may be more beneficial to focus on other variables when developing strategies.

2) Monthly income, buffalo milk protein content, time of day, and level of sleep had a greater effect on the frequency of buffalo milk purchases.

3) There was a significant positive correlation between monthly income, protein content of buffalo milk, time of day and frequency of purchase. Consumers with higher incomes tend to purchase buffalo milk more frequently, which may be due to their increased purchasing power and possibly greater health consciousness. In addition, the protein content of buffalo milk appears to be a strong selling point, suggesting that consumers value the nutritional value it provides. Therefore, publicizing the protein-rich nature of buffalo milk may be a key strategy to increase sales.

4) There is a negative correlation between the effect of sleep quality and the frequency of purchasing buffalo milk, consumers with poorer sleep quality on the contrary increase the frequency of purchasing buffalo milk, consumers with better sleep quality tend to drink buffalo milk less often, so we can assume that buffalo milk contributes to sleep quality to some extent.

4. Conclusions and recommendations

4.1. Conclusions

In this study, a questionnaire was issued to Shanghai's Generation Z population through the network survey method and literature survey method. After the questionnaire was issued, K-means clustering was utilized to cluster different consumers, and at the same time, the consumption cognition and consumption characteristics of buffalo milk of different consumer groups were analyzed; after the characteristics of the population were obtained, the influencing factors on the frequency of buffalo milk purchasing were further analyzed based on the logistic model. After obtaining the characteristics of the population, we further analyzed the factors influencing the

purchase frequency of buffalo milk based on the logistic model, and finally obtained the following conclusions:

1) The market for buffalo milk is characterized by high production costs, low willingness to buy among low-income groups, and not very high visibility as a consumer barrier.

2) With high demand from Asian countries and the nutritional value of buffalo milk being explored, coupled with the broadening of purchasing channels such as online group-buying platforms, the market opportunity for buffalo milk is significant.

4.2. Recommendations

Combining the above findings with the research in this paper we can give the following recommendations:

1) Optimize the taste and brand, while launching products with better taste and more in line with consumer demand, optimize the brand image and brand promotion strategy to improve brand awareness and reputation and increase market share.

2) Through scientific production process and technology, buffalo milk is made to provide richer nutritious proteins and long-lasting satiety, while emphasizing its difference from other milks and reinforcing its functions in terms of health and nutrition. Aiming at female consumers, we offer buffalo milk products with beauty benefits.

3) Buffalo milk companies should utilize social media to provide more information, information and communication interaction in their promotion. Establish social media accounts to post excellent information and effective marketing so that consumers can learn more about their products. Also, attract new market share with advertising and branding on social media.

4) Buffalo milk companies should adopt competitive and appropriate pricing strategies to give them an edge in market share. In this way, the company can survive and thrive in the fierce competition. Advantage through excellent taste, nutritional value (protein content, etc.) and branding.

References

- [1] JIANG Handong, PENG Xiaofeng, ZHANG Huazhi et al. Development of dairy buffalo industry in Lingshan County [J]. China Animal Husbandry, 2023 (08): 27 - 29.
- [2] LIU Yanjing, ZHU Jing, LIU Chang. On the development path of buffalo milk industry in Guangxi[J]. China Dairy Cattle, 2021 (07): 61 - 64.DOI: 10.19305/j.cnki.11 - 3009/s.2021.07.011.
- [3] ZHAO Ye, LI Yuan, ZHAO Cunzhao et al. Development of buffalo milk cream ice cream [J]. China Dairy Industry, 2021, 49 (03): 60 - 64.DOI: 10.19827/j. issn1001 - 2230.2021.03.012.
- [4] FANG Wenyuan, TANG Qingfeng, ZHU Geying et al. Measures to promote the quality and efficiency of southern specialty dairy buffalo industry [J]. Contemporary animal husbandry, 2020 (07): 60 - 62.
- [5] WANG Jinmei, YANG Yuan, MIAO Yongwang. Research on the development status and countermeasures of dairy buffalo industry in China [J]. Chinese Journal of Animal Husbandry, 2020, 56 (02): 177 - 180.DOI: 10.19556/j.0258 - 7033.20190625 - 06.
- [6] HUANG Jiaxiang, HUANG Feng, QIN Guangsheng et al. An overview of the development of China's dairy buffalo industry - country profile and scientific and technological innovation [J]. China Dairy Cattle, 2019 (11): 1 - 8.DOI: 10.19305/j.cnki.11 - 3009/s.2019.11.001.
- [7] WANG Jinmei, YANG Yuan, MIAO Yongwang. Dilemma and way out of Yunnan buffalo milk industry development [J]. Heilongjiang Animal Husbandry and Veterinary Medicine, 2019 (06): 11 - 14.DOI: 10.13881/j.cnki.hljxmsy.2018.03.0483.
- [8] WANG Jinmei, YANG Yuan, MIAO Yongwang. Inspiration of the development of buffalo milk industry in India and Italy to China [J]. China dairy industry, 2018, 46 (10): 32 - 35.
- [9] Yao M, He P.C. Research on the initial value selection problem of K-means algorithm [J]. Fujian Computer, 2023, 39 (07): 57 - 61.DOI: 10.16707/j.cnki.fjpc.2023.07.011.
- [10] Xinyu Zhu. Creative Design of Milk Tea Flash Shop [J]. Beauty and Times (City Edition), 2023 (04): 143.