Pure Mathematics 理论数学, 2025, 15(10), 187-196
Published Online October 2025 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2025.1510262

Hans汉斯

基于GMM算法的膨化食品消费者画像研究

孙翠雪[1]，牟唯嫣[1]，吕　磊[2]

[1]北京建筑大学理学院，北京
[2]南京电子设备研究所，江苏　南京

收稿日期：2025年9月17日；录用日期：2025年10月20日；发布日期：2025年10月30日

摘　要

本研究通过收集问卷数据，基于GMM算法构建膨化食品消费者画像。首先通过探索性因子分析(EFA)对数据进行降维，结合Spearman相关性检验筛选核心变量，运用GMM聚类分析，最终将膨化食品消费者划分四类。高消费潜力型青少年具有单一社交消费属性，偏好咸口与复合口味。多维度灵活型社交场景消费显著，购物方式灵活，偏好复合口味且价格选择单一。口味价格聚焦型计划消费性强，多渠道购买，口味与价格选择均单一。价格敏感型消费能力弱，依赖促销驱动。

关键词

GMM算法，消费者画像

A Study on Consumer Profiling for Puffed Snacks Based on the GMM Algorithm

Cuixue Sun[1], Weiyan Mu[1], Lei Lyu[2]

[1]School of Science, Beijing University of Civil Engineering and Architecture, Beijing
[2]Nanjing Electronic Equipment Research Institute, Nanjing Jiangsu

Received: September 17, 2025; accepted: October 20, 2025; published: October 30, 2025

Abstract

This study collects questionnaire data and constructs a portrait of puffed food consumers based on the GMM algorithm. Firstly, exploratory factor analysis (EFA) is used to reduce the dimensionality of the data, and core variables are screened by combining with the Spearman correlation test. Finally, through GMM cluster analysis, puffed food consumers are divided into four categories. Teenagers with high consumption potential have a single social consumption attribute and prefer salty and compound flavors. The multi-dimensional flexible type shows prominent consumption in social scenarios, with flexible shopping methods, a preference for compound flavors, and a single price

文章引用: 孙翠雪, 牟唯嫣, 吕磊. 基于 GMM 算法的膨化食品消费者画像研究[J]. 理论数学, 2025, 15(10): 187-196.
DOI: 10.12677/pm.2025.1510262

choice. The taste-price focused type has strong planned consumption behavior, adopts multi-channel purchasing, and shows single choices in both taste and price. The price-sensitive type has weak consumption capacity and relies on promotions to drive purchases.

Keywords

GMM Algorithm, Consumer Profiling

1. 引言

膨化食品作为休闲零食市场的重要组成部分，近年来在中国消费规模持续扩大，消费群体结构呈现年轻化与需求多元化的显著特征[1]。随着行业竞争加剧与市场细分程度提升，如何精准识别不同类别消费者的行为模式与偏好特征，成为企业优化产品结构、制定精准营销策略的关键。传统的消费者细分方法多依赖于人口统计学变量或单一维度的消费行为指标，难以有效捕捉多因素交互作用下形成的复杂群体结构，尤其在高维、非线性关系的数据环境中，其解释力与预测精度均存在一定局限。

2. 问卷基本特征分析

2.1. 问卷设计及信效度检验

本文通过线上问卷针对受访者购买过膨化食品、购买过狗牙儿产品和未购买过膨化食品的三种不同现状设置了不同角度的问题，调查消费者的消费偏好、购买行为等相关信息，以实现我们的调查目的。通过预调查，测算出合理的问卷数量需要大于 278 份。回收有效问卷数目共 382 份，其中有 330 份问卷购买过膨化食品，认为问卷样本量合理。对问卷进行信效度检验得到表 1、表 2。

由表 1 可知，前两个量标题克隆巴赫 Alpha 系数均大于 0.80，量表信度良好，具有较高的一致性。但第三个量表的各个选项克隆巴赫 Alpha 系数在 0.5~0.6 之间，量表整体的克隆巴赫 Alpha 系数为 0.665。虽未达到非常理想的信度水平(大于 0.7)，但仍具有一定的内部一致性。由表 2 可知，前两个题项的变量相关性较高，适合进行因子分析，其巴特利特球形度检验显著性均小于 0.05，表明变量间存在显著相关性，满足因子分析前提。题项 3 量表 KMO 值为 0.680，虽相对略低，但也可进行因子分析，显著性小于 0.05。最后，所有量表题综合检验时，KMO 值为 0.907，显著性小于 0.05，意味着整体数据适合进行因子分析，能够进一步深入探究问卷量表的结构效度，为后续研究提供有力支撑。

Table 1. Questionnaire reliability analysis table
表 1. 问卷信度分析表

题项	克隆巴赫 Alpha	基于标准化项的克隆巴赫 Alpha	项数
各因素对选择膨化食品的影响度	0.899	0.898	8
狗牙儿食品的包装设计评分	0.869	0.869	5
狗牙儿品牌与其他品牌相比优势	0.665	0.664	4
综合	0.88	0.882	17

Table 2. Questionnaire validity analysis table
表 2. 问卷效度分析表

题项	KMO 取样适切性量数	巴特利特球形度检验显著性
各因素对选择膨化食品的影响度	0.930	0.000
狗牙儿食品的包装设计评分	0.874	0.000
狗牙儿品牌与其他品牌相比优势	0.680	0.000
综合	0.907	0.000

2.2. 关键变量的描述统计

2.2.1. 受访者性别、年龄、所在年级分布

在调查回收的有效问卷中，女性比例为 52.62%，男性比例为 47.38%，较为符合真实情况，验证本次调查的随机性。

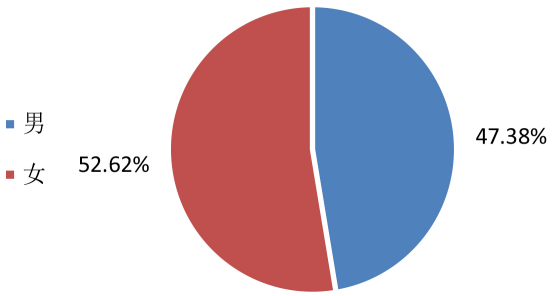

Figure 1. Gender pie chart
图 1. 性别饼状图

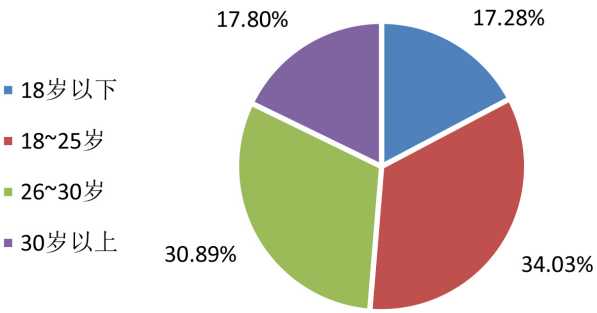

Figure 2. Gender and age pie chart
图 2. 年龄饼状图

由图 1、图 2 可知，本次受访者的年龄分布集中分布在 30 岁以下，是膨化食品的重点针对的消费者市场，占总样本的 82.20%。其中 18~25 岁受访者最多，占总样本的 34.03%，其次是 26~30 岁的受访者，占总样本的 30.89%。

另外，根据受访者所在年级分布可以看出(图 3)，本次调查的受访者中较多的是硕士研究生和本科生，分别占所有受访者的 34.82%和 25.39%。此外，博士研究生占比 16.75%，高中及以下学生占比 13.61%，“其他”的受访者占比最少，仅有 9.42%。受访者以在校学生为主，符合本报告对调查对象的预期。

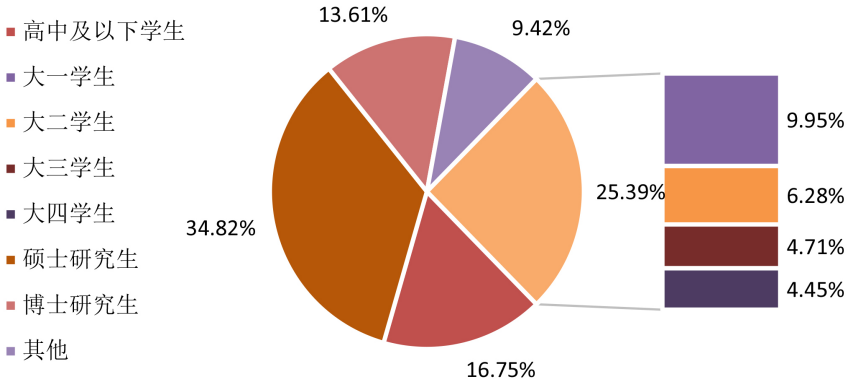

Figure 3. Pie chart of grade distribution among respondents
图 3. 受访者所在年级分布饼状图

2.2.2. 受访者所在地区及所在城市定位分布

从所在地区的分布可以看出(图 4)，本次调查的受访者在华东地区的分布最多，占所有受访者的27.75%。东北、华北、西南地区占比相近，分别占 17.02%、14.14%、13.87%。华中、西北地区都占所有受访者的 10%左右。华南地区受访者人数最少，仅占 6.81%。

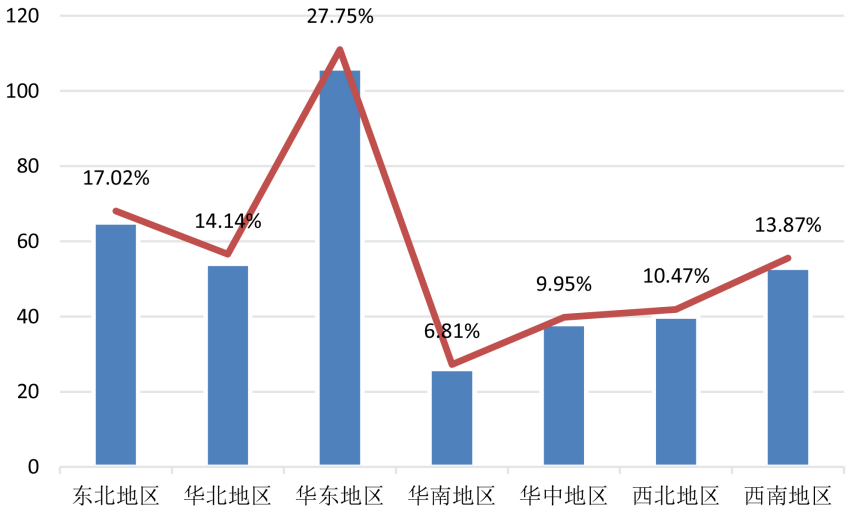

Figure 4. Geographic distribution of respondents
图 4. 受访者所在地区分布

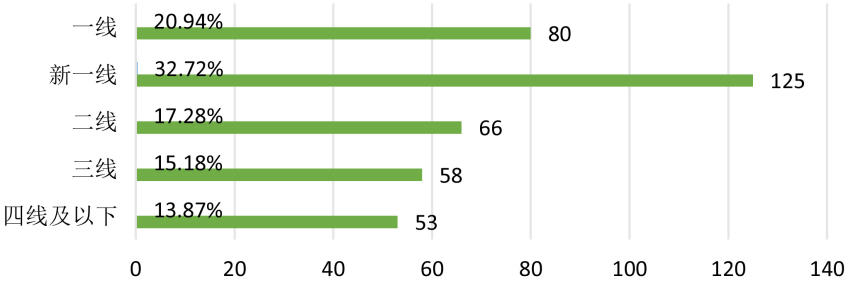

Figure 5. Comprehensive development level of the city where respondents reside
图 5. 受访者所在城市发展综合水平

从受访者所在城市的定位分布中可以看出(图 5)，本次调查的受访者大部分位于经济情况相对较好的城市，其中位于新一线城市的受访者最多，占所有受访者的 33%，一线城市的受访者次之，占所有受访者的 21%。位于二线、三线、四线及以下城市的受访者分布相近，分别占所有受访者的 17%、15%、14%，表明低线城市消费者在膨化食品市场中也占有一席之地。本次调查的主要对象为学生群体，故而所在城市定位分布与一般大学所在城市定位分布接近，表明该样本对于研究大学生膨化食品市场情况较为合理。

2.2.3. 受访者月可支配收入分布

图 6 展示了本次调研中受访者月可支配收入的分布情况，通过柱状图和正态分布曲线可以看出，收入在 2001~3000 元和 1001~2000 元的受访者占比较高，1000 元及以下和 3000 元以上的受访者数量相对较少，整体呈现出中间高、两侧低的分布形态，呈近似正态分布。这种收入分布表明膨化食品市场的主要消费群体集中在有一定消费能力的中等收入人群。

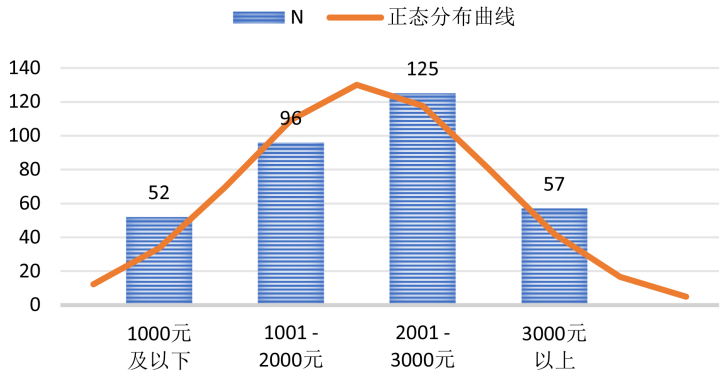

Figure 6. Distribution of respondents' disposable income
图 6. 受访者可支配收入分布图

3. 数据处理

3.1. 数据降维

针对五个多选题的高维度特征，分别采用探索性因子分析[2] (EFA)进行降维处理，每个多选题提取两个公共因子，结果如表 3 所示。

结果显示，各量表均通过 Bartlett 球形度检验($p < 0.001$)，其中品类偏好量表 KMO 值最高(0.665)，价格接受度量表 KMO 值最低(0.443)。通过主成分分析法提取两个公因子，各量表累积方差贡献率介于 61.49%~79.94%之间：消费场景偏好量表提取日常休闲场景与社交聚会场景两个因子，品类偏好量表分离出经典酥脆零食与多元膨化零食维度，口味偏好量表形成咸口偏好与复合口味偏好因子，价格接受度量表识别出价格敏感型与价格中性型两类消费者，购买渠道量表区分即时性渠道与计划性渠道选择。所有分析均采用方差最大化正交旋转，确保因子结构清晰可解释，为后续聚类模型提供变量支持。

3.2. 相关性分析

基于高斯混合模型(GMM)的聚类分析需求，本研究首先对变量进行筛选：针对上节提取的 10 个核心变量与问卷基础属性维度、消费习惯维度共 18 个变量开展 Spearman 秩相关性检验[3]，剔除与其他变量高度相关的冗余变量和几乎不相关的无关变量。最终提取年龄、所在城市定位、每月可支配金额、聚会外出场景、多元膨化零食、咸口偏好、复合口味偏好、价格中性型倾向、计划性渠道选择 9 个特征变量作为聚类分析指标，确保变量间独立性与数据信息完整性。

Table 3. Multi-choice item dimension reduction results table
表 3. 多选题项降维结果表

问题	KMO 值	巴特利特球形度检验	累积旋转载荷平方和	提取的公因子	包含变量
您通常在什么场景下食用膨化食品？	0.523	<0.001	62.239%	日常休闲场景	工作学习和休闲放松
				聚会外出场景	聚会活动和外出旅行
您更倾向于选择以下哪些膨化食品？	0.665	<0.001	61.487%	经典酥脆零食	薯片/蚕豆/比萨卷
				多元膨化零食	虾片、雪米饼等其他种类
您更倾向于选择以下哪些口味的膨化食品？	0.581	<0.001	79.943%	咸口类	重口类和轻咸类
				复合口味	甜口类、清新类和猎奇类
您更倾向于以下哪个膨化食品的价格区间？	0.443	<0.001	75.698%	价格敏感型	6 元以下及 20 元以上
				价格中性型	6~10 元、11~15 元和 16~20 元
您通常购买膨化食品的渠道是？	0.569	<0.001	64.409%	即时性渠道	校内超市、校外便利店和社区团购
				计划性渠道	大型超市、网店及直播间

4. GMM 算法

本研究选择高斯混合模型(GMM) [4]作为聚类分析方法，主要基于其处理复杂数据结构和提供丰富解释的独特优势。传统的硬聚类算法(如 K-Means)虽然计算高效，但其存在两点主要局限：一是假设聚类呈球形分布且大小均匀，难以捕捉现实数据中复杂的几何形状；二是采用“非此即彼”的硬分配原则，强制将每个数据点唯一地归属于一个簇，这在处理具有混合特征或边界模糊的个体时显得过于僵化。GMM 算法作为一种基于概率模型的软聚类算法[5]，非常适合该问卷涉及的消费者消费偏好与购买行为部分的数据，因为这部分数据具有不确定性，例如，一个消费者可能同时兼具“价格敏感”和“品牌忠诚”的混合特征，因此本文选择采用 GMM 算法作为聚类分析方法。

4.1. GMM 模型定理

GMM 核心定理是其概率密度函数[6]，假设数据由 K 个高斯分布混合生成，则其概率密度函数为：

$$P(x)=\sum_{k=1}^{K}P(c_k)P(x|c_k)=\sum_{k=1}^{K}\pi_k N(x|\mu_k,\Sigma_k)$$

其中，x 为一组多维数据点；$P(x|c_k)\sim N(x|\mu_k,\Sigma_k)$ 服从高斯分布；K 是混合成分的数量，即潜在高斯簇的个数；π_k 是第 k 个高斯分量的混合权重，满足 $\sum_{k=1}^{K}\pi_k=1$，是各高斯分布在整个模型中的权重；$N(x|\mu_k,\Sigma_k)$ 是第 k 个高斯分布的概率密度函数，对应均值向量 μ_k 和协方差矩阵 Σ_k。每个数据点以一定的概率属于各个高斯分量，使得 GMM 能够捕捉数据分布的复杂形态。

4.2. EM 算法原理

高斯混合模型的学习通常采用期望最大化(Expectation-Maximization, EM)算法，这是一种迭代优化算法，旨在最大化观测数据的对数似然函数[7]。EM 算法分为两步：

(1) E 步骤(期望步骤)

在给定的多维高斯分布下，计算每个数据点 x_i 属于每个高斯分量 k 的后验概率(属于各个类别的概率)：

$$P(c_k \mid x_i) = \frac{\pi_k N(x_i \mid \mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k N(x_i \mid \mu_k, \Sigma_k)}$$

(2) M 步骤(最大化步骤)

在 M 步骤中，利用 E 步骤计算得到的概率，更新模型参数以最大化期望对数似然函数。

$$N_k = \sum_{i=1}^{N} P(c_k \mid x_i)$$

① 更新混合权重

$$\pi_k = \frac{N_k}{N}$$

② 更新均值向量

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} P(c_k \mid x_i) x_i$$

③ 更新协方差矩阵

重复 E 步和 M 步直至收敛或达到最大迭代次数。

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} P(c_k \mid x_i)(x_i - \mu_k)^{\mathrm{T}}(x_i - \mu_k)$$

4.3. GMM 模型应用

基于高斯混合模型(GMM)的聚类分析提取 4 个高斯分量作为聚类中心，实现消费者群体的四维聚类划分。采用 t-SNE 降维技术将高维聚类结果映射至二维空间进行可视化展示，如图 7 所示。四种颜色对应了四类消费者群体的投影分布。聚类中心在二维空间中的位置清晰反映了类间距离与密度差异。各类的边界较为清晰，说明分类效果较好。

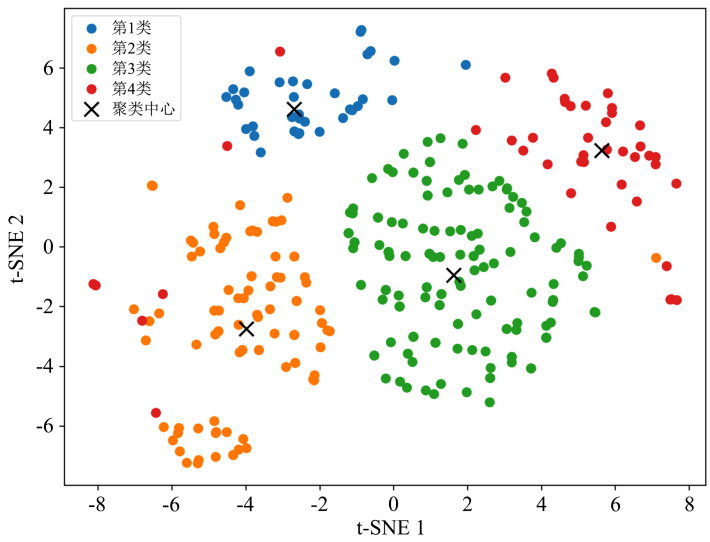

Figure 7. GMM clustering results (after t-SNE dimension reduction)
图 7. GMM 聚类结果(t-SNE 降维后)

5. 用户画像呈现

基于 GMM 聚类所得结果，将消费者划分为高消费潜力型青少年、多维度灵活型消费者、口味价格聚焦型消费者以及价格敏感型消费者，并绘制用户画像(图 8、图 9、图 10 以及图 11)。

高消费潜力型青少年：目标群体为 18 岁以下青少年，主要分布于一线城市，月均可支配金额超过 3000 元，属于高消费人群。该群体消费场景呈现显著单一社交属性，消费渠道呈现计划性特征，通常提前制定采购清单并系统性整合促销信息，形成以网络平台与大型商超为核心的双线消费模式。品类选择集中于虾片、雪米饼等七类零食中的四类，呈现咸口与复合口味双偏好特征且价格敏感度呈现中性特征。

Figure 8. Profile of high-potential youth consumers
图 8. 高消费潜力型青少年画像

多维度灵活型消费者：目标群体为 18~30 岁青年，主要分布于经济发达地区，月均消费 2000~3000 元。该群体消费场景具有显著社交属性，倾向于在聚会、旅行等场景中消费膨化食品。购买呈现计划性，根据场景需求灵活切换购物方式。品类选择集中于虾片、雪米饼等七类零食中的四类，对复合口味产品表现出明显偏好并在价格选择方面呈现单一化。

Figure 9. Multi-dimensional flexible consumer profile
图 9. 多维度灵活型消费者画像

口味价格聚焦型消费者：目标群体为 18~30 岁青年，主要分布于一线/新一线城市，可支配金额 1000~3000 元，具备一定消费能力。消费行为呈现显著社交属性，偏向在社交场景中消费膨化食品。消费决策呈现显著计划性特征，形成线上线下多渠道覆盖的消费模式。品类选择集中于虾片、雪米饼等七类零食中的四类，在膨化食品口味和价格选择上均呈现单一化。

Figure 10. Profile of taste-and-price-focused consumers
图 10. 口味价格聚焦型消费者画像

价格敏感型消费者：以 25 岁以下青年为主，主要分布于四线及以下城市，月均可支配收入低于 2000 元，消费能力不高。受限于经济水平，该群体呈现明显消费特征：社交活动参与度较低，消费渠道集中于单一平台或场景，对零食品类需求呈现高度聚焦性，且价格敏感度极高。因此促销活动是驱动该群体产生购买行为的核心驱动力，其消费决策高度依赖折扣力度与限时优惠。

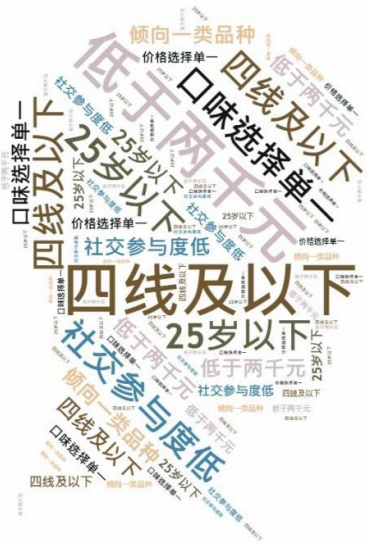

Figure 11. Price-sensitive consumer profile
图 11. 价格敏感型消费者画像

6. 结论

1. 地域消费分层显著体现在不同城市层级的消费能力差异上：一线城市存在高消费青少年群体(月均超 3000 元)与低消费青年群体(月均不足 2000 元)的二元结构；新一线城市青年群体呈现中等消费能力(月均 2000~3000 元)；低线城市青年群体消费能力大多较低(月均不足 1000 元)。

2. 社交场景对消费行为具有普遍驱动作用，所有群体均在聚会、旅行等社交场景中高频消费膨化食品。高消费群体表现出场景单一性特征，低线城市群体则兼具社交与日常场景消费需求，呈现双场景活跃特征。

3. 消费渠道选择与消费能力呈现强相关性：高消费群体采用“大型商超 + 网络平台”的双线计划性采购模式；低消费群体消费渠道集中于单一平台或场景，且价格敏感型群体的消费决策高度依赖促销活动。

4. 价格敏感度与消费能力呈负相关关系：月均消费 2000 元以下群体价格敏感度极高，促销活动是驱动其购买行为的核心因素；高消费群体则呈现价格中性特征，更注重产品品质与消费场景的适配性。

参考文献

[1] 中国报告大厅网. 2024 年膨化食品行业现状分析: 膨化食品企业加大产品创新力度[EB/OL]. 2024-07-22. https://www.chinabgao.com/freereport/94967.html, 2024-12-12.

[2] 黄文林, 李政. 企业战略与市场分析研究——基于探索性因子分析与贝叶斯网络模型构建[J]. 中国经贸导刊, 2025(6): 139-141.

[3] 茆诗松, 程依明, 濮晓龙. 概率论与数理统计[M]. 北京: 高等教育出版社, 2019.

[4] 李婧. 基于 GMM 的 EM 优化算法的应用与研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工程大学, 2018.

[5] 姚博凡, 邓红平, 蔡铭. 基于随机抽样 GMM 的城市交通运行状态模式分类[J]. 计算机工程, 2020, 46(12): 36-42.

[6] 张美霞, 李丽, 杨秀, 等. 基于高斯混合模型聚类和多维尺度分析的负荷分类方法[J]. 电网技术, 2020, 44(11): 4283-4296.

[7] 茆诗松, 程依明, 濮晓龙. 概率论与数理统计教程[M]. 北京: 高等教育出版社, 2019.