**一、课题综述**

* 1. **课题说明**

实验分工如下：

* 1. **课题目标**

本研究旨在利用新媒体用户画像数据集，通过多种聚类算法（如K-means、DBSCAN、层次聚类、GMM聚类）对用户进行细分，以揭示不同用户群体的行为特征和偏好，并探讨聚类结果在新媒体平台中的应用价值，如个性化推荐系统优化、广告精准投放、用户行为预测等，为平台运营和营销策略提供科学依据。

数据集链接：https://www.kaggle.com/datasets/jojoyin/new-media-user-analysis-dataset/data?select=5-.csv

* 1. **课题数据集**

数据集来源于Kaggle，包含用户的生活习惯和消费行为的多项调查问卷结果。每个用户对多个问题进行了评分，评分范围从1到5，其中1表示“非常不同意”，5表示“非常同意”。数据集共有676个样本（用户），每个样本有12个特征。部分特征如下：

* 我尽量每天都定时入睡：用户是否努力保持每天定时入睡的习惯。
* 我每天三餐时间一致：用户是否每天按时吃三餐。
* 我不会随意打破作息时间：用户是否严格遵守自己的作息时间。
* 对我来说品质比价格重要：用户在购买商品时更看重品质还是价格。
* 外出时候宾馆需要很干净：用户对外出住宿环境的清洁程度的要求。
* 我的衣服等日用品都是品牌的：用户是否倾向于购买品牌商品。
* 我每周都会运动：用户是否定期进行体育锻炼。
* 我很少暴饮暴食：用户是否经常暴饮暴食。
* 我早睡早起：用户是否有早睡早起的习惯。
* 关注可穿戴设备：用户是否对可穿戴设备感兴趣。
* 新的东西出来想尝试：用户是否喜欢尝试新事物。
* 就算新品贵一些也会买：用户是否愿意为新产品支付更高的价格。
* 年龄
* 性别

**二、实验报告设计**

* 1. **数据准备**

**2.1.1 数据集获取与数据加载**

数据集于2024年2月在[Kaggle](https://www.kaggle.com/datasets/taweilo/taiwan-air-quality-data-20162024" \t "_new)发布，使用pandas库将其以CSV形式加载到Python3环境

**2.1.2 数据探索性可视化**

* 1. **数据预处理**

**2.2.1 初步探索**

数据集无缺失值、重复值、异常值，无需处理缺失值，无需删除重复值，无需剔除异常样本；分离id列

**2.2.2 特征归一化**

使用最小最大归一化将数值型特征（年龄）归一化

**2.2.3 因子分析降维**

* **降维原因：**

因子分析可避免高维数据导致的“维度灾难”，即随着维度增加，数据点间的距离变得非常大，使聚类算法难以找到有意义的簇；还可增强模型解释性，因子分析得到的因子通常可分析出其实际意义，基于因子的聚类的每个簇可以用少数几个因子来描述，而不是大量的原始特征；还可以去除原始数据多重共线性。

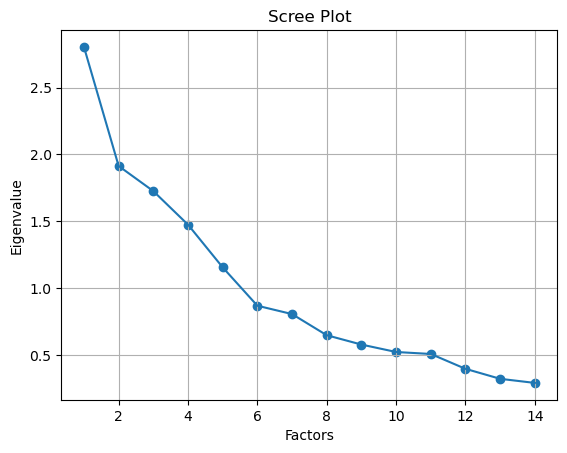
* **步骤及原理：**
* KMO和巴特利特球形检验：

上述预处理后数据集检验结果如下，说明数据适合进行因子分析

|  |  |
| --- | --- |
| 指标 | 结果 |
| KMO测度 | 0.6701 |
| 巴特利统计量 | 2158.6463 |
| p值 | 0.0001 |

* 根据Kaiser准则，保留特征值大于1的因子，保留的因子数量为4

特征值表示每个因子解释的方差量。特征值越大说明该因子解释的数据变异越多，而等于1表示该因子解释的方差量等于一个变量的平均方差，因此特征值大于1的因子被认为具有显著解释能力。



* 因子载荷表与各因子解释

分析后得到因子载荷表如下：表中的数值表示每个变量在相应因子上的权重，绝对值越大表示该变量与该因子的关系越强



根据各变量与各因子相关性，因子含义可解释如下：

* Factor1：健康生活习惯（7~9）
* Factor2：注重品质（4~6）
* Factor3：创新与消费倾向（10~12）
* Factor4：规律作息（1~3）

注：后续实验发现年龄和性别会引入过多噪声，影响聚类效果，故删除

* **最终用于聚类的数据表：**



* 1. **模型选择**

**2.3.1 模型评估指标**

由于没有先验知识（数据集没有标签），本次实验使用以下3种不需要真实标签的内部指标：

* **Silhouette Coefficient（轮廓系数）**

轮廓系数衡量的是样本点与其所在簇内其他点的平均距离（a）与该样本点与其他最近簇中所有点的平均距离（b）之间的关系。

对于每个样本点i：

* a(i) 是样本点i到其所在簇内其他点的平均距离。
* b(i) 是样本点i到最近的一个不同簇的所有点的平均距离。

计算公式为：

s(i) 的取值范围是[-1, 1]，值越大表示聚类效果越好

* **Calinski-Harabasz Index（CH指数）**

CH指数通过比较簇间散布矩阵和簇内散布矩阵评估聚类的效果。

计算公式为：

其中,B表示簇间的散布矩阵，W表示簇内的散布矩阵，k是簇的数量，n是总样本数。CH指数越高，说明簇间的分离度好，簇内的紧凑性好。

* **Davies-Bouldin Index（DB指数）**

DB指数基于簇内相似性和簇间差异性的比率来评估聚类效果。

计算公式为：

其中*k* 是簇的数量，*σi*​ 和*σj*​ 分别是第i个和第j个簇的平均距离，*d*(*ci*​,*cj*​) 是两个簇中心的距离。DB指数越低，说明聚类效果越好。

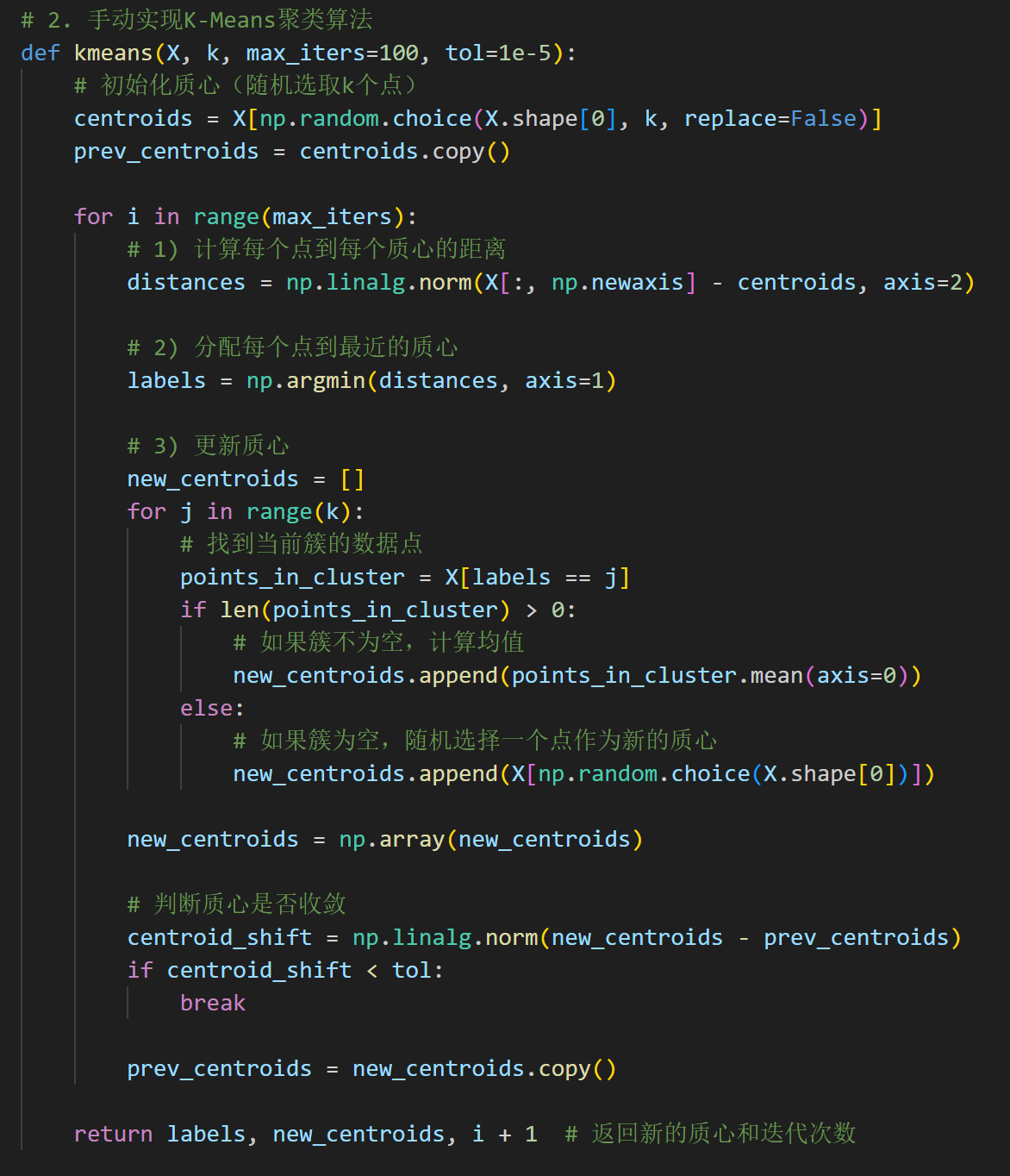
**2.3.2 K-means聚类**

K-means聚类是一种流行且简单的无监督机器学习算法，用于将数据集划分为几个簇，使簇内的数据尽可能相似，不同簇之间尽可能相异。核心思想是通过迭代过程最小化簇内所有点与该簇中心点之间的平方误差之和。本次实验通过手写实现和调库实现两种方式进行kmeans聚类并解释其结果。

基本步骤如下：

* **初始化：**首先需要指定簇的数量𝑘，然后随机选择𝑘个数据点作为初始簇中心（centroid）。也可以采用一些启发式方法来选择初始中心点，比如K-means++。（这也是之后每次聚类结果不同的原因，下述采用一次典型结果做分析）
* **分配：**对于数据集中的每一个点，计算它与各个簇中心之间的距离，并将其分配给距离最近的那个簇。
* **更新：**重新计算每个簇的中心点位置，是该簇中所有点的平均值。
* **迭代：**重复执行“分配”和“更新”步骤，直到满足停止条件。停止条件包括达到最大迭代次数、所有点的簇分配不再改变或者簇中心的变化小于某个阈值。

核心函数实现如下：



**2.3.3 DBSCAN聚类**

DBSCAN是一种基于密度的聚类算法，能够发现任意形状的簇，并且能够识别噪声点。与传统的基于距离的聚类算法（如KMeans）不同，DBSCAN不需要预先指定簇的数量，而是根据数据点的密度来动态确定簇的边界。

* 核心点(Core Point)：如果一个点邻域（半径eps）内的点数大于等于阈值min\_samples，则称核心点。
* 边界点(Border Point)：一个点不是核心点但在某核心点邻域内，则称边界点。
* 噪声点 (Noise Point)：既不是核心点也不是边界点的点。

核心函数实现如下：



**2.3.4 层次聚类**

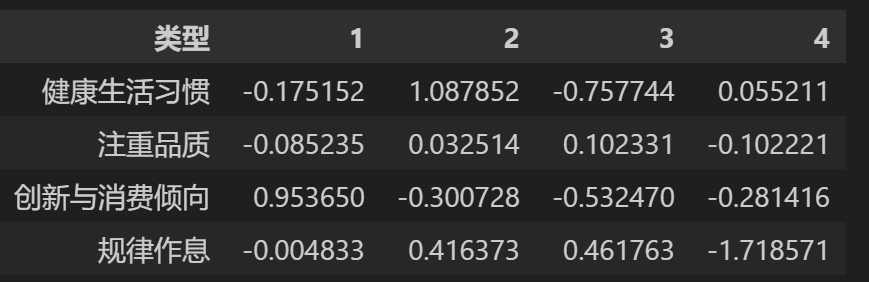
* 1. **模型搭建**

**2.4.1 K-means聚类**

首先通过计算不同聚类数目下的轮廓系数（Silhouette Score）评估聚类效果，自动选择最佳聚类数目，过程可视化如下：

****

* 当选择聚类簇数为4时，本次相应结果如下：

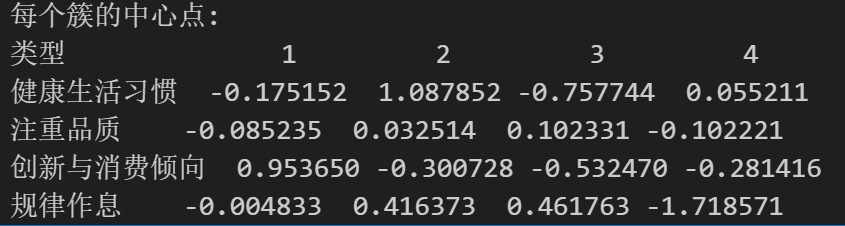
****

上述k-means聚类将用户分为四类，每类用户特征含义分析见2.5.1

* 计算相应评价指标如下：

|  |  |  |
| --- | --- | --- |
| Silhouette Coefficient | Calinski-Harabasz Score | Davies-Bouldin Score |
| 0.4009 | 151.1759 | 1.4152 |

* 各簇中心点：

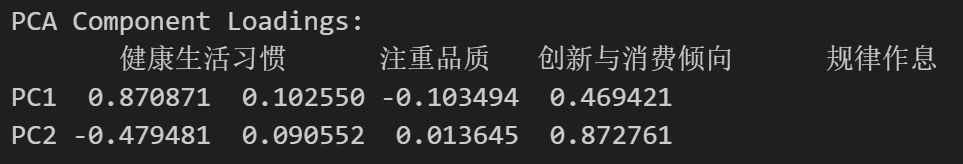
****

**2.4.2 DBSCAN聚类**

由于其基于密度的特性，而在高维数据中，密度的概念会变得模糊，DBSCAN的性能在因子分析后的数据上进行聚类的效果不好，无论怎样扩大搜索半径，减少同簇样本数，都只能聚成一类：

考虑使用PCA方法对所得因子特征再次降维，降至2维。

降维后主成分载荷矩阵如下：



使用降维后的二维数据进行DBSCAN聚类，经过多组参数选择，最好效果的（半径eps=0.5，min\_samples=6）结果如下：

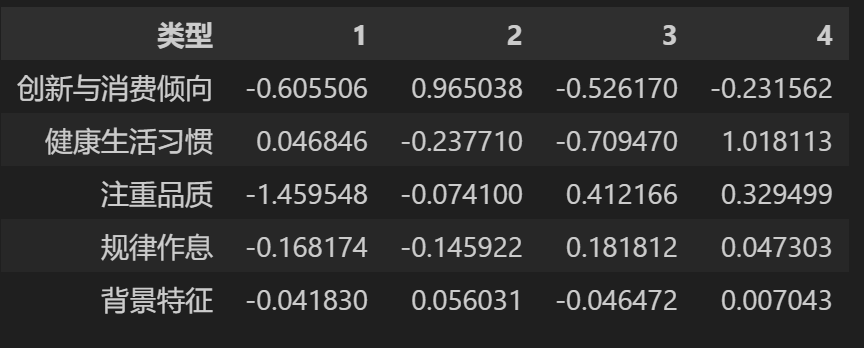
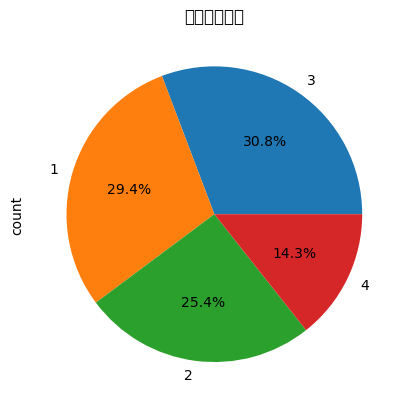
|  |  |  |
| --- | --- | --- |
| Silhouette Coefficient | Calinski-Harabasz Score | Davies-Bouldin Score |
| 0.0237 | 0.4402 | 55.8094 |

**2.4.3 层次聚类**

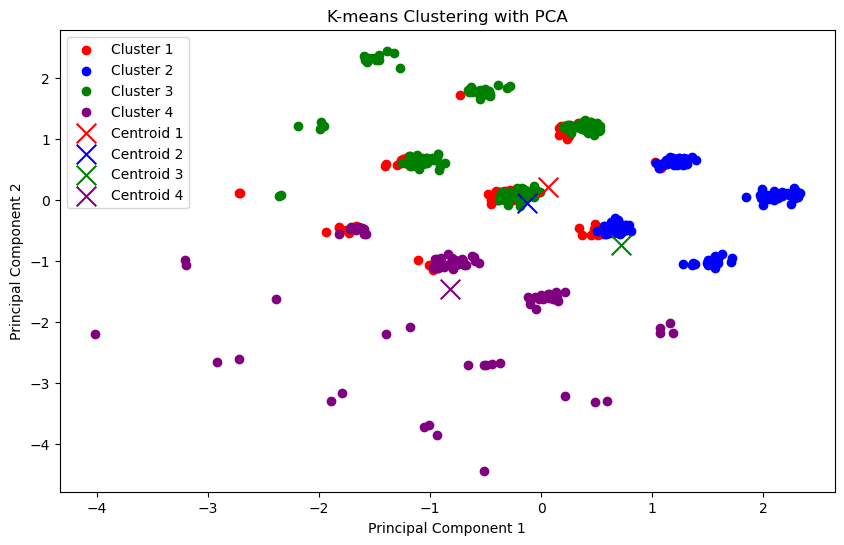
* 1. **结果分析与可视化**

各模型聚类效果可视化如下：

* + 1. **K-means结果分析与可视化**
* 聚类结果与各类别人数占比：

* PCA降至2维后聚类可视化：



* 上述k-means聚类将用户分为四类，每类用户特征分析如下：

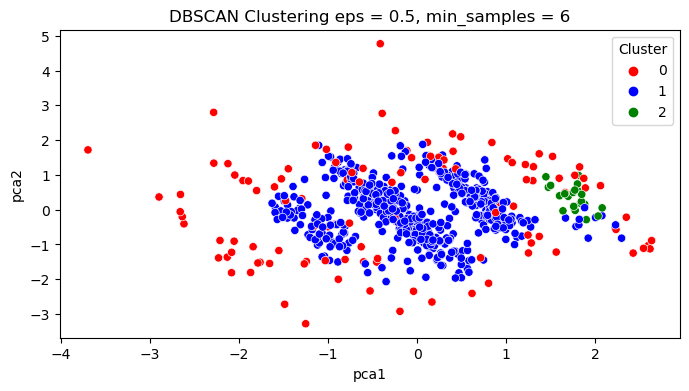
|  |  |
| --- | --- |
| **用户类型** | **用户特征** |
| **类型一** | 创新意识和消费倾向高的用户，可能更愿意尝试新事物和新产品 |
| **类型二** | 注重健康生活习惯和生活规律的用户，对创新和消费的兴趣较低，略重品质 |
| **类型三** | 生活作息有规律但不太注重健康生活习惯的用户，对创新和消费的兴趣较低，略重品质 |
| **类型四** | 生活作息不规律，对健康生活习惯和创新消费的兴趣一般 |

* 评价指标

|  |  |  |
| --- | --- | --- |
| Silhouette Coefficient | Calinski-Harabasz Score | Davies-Bouldin Score |
| 0.2009 | 151.1759 | 1.4152 |

轮廓系数0.2009的表明聚类效果中等，大部分样本点与其所在簇内的其他点较接近，与其他簇的距离相对较远，不过还有改进空间；CH指数 151.1759很高，说明当前的聚类结果在簇间分离和簇内紧凑性方面表现良好；DB指数值为1.4152，相对较低，表明聚类结果的簇间差异较大，簇内相似性较高，效果较好。

* + 1. **DBSCAN结果分析与可视化**
* 聚类结果（经2.4.2处理的两个主成分上）：



* 评价指标

|  |  |  |
| --- | --- | --- |
| Silhouette Coefficient | Calinski-Harabasz Score | Davies-Bouldin Score |
| 0.0237 | 0.4402 | 55.8094 |

结果显示Silhouette Coefficient (轮廓系数) 接近0，表示样本与其同一簇中的其他样本的距离和与其他簇中的样本的距离相近，聚类效果一般；CH-Score值较小，DB-Score值较大，表明聚类效果不甚理想。

分析原因：由于DBSCAN基于密度的特性，而在高维数据中，密度的概念会变得模糊，DBSCAN的性能在因子分析后的数据上进行聚类的效果不好，在原始数据集上同理，无论怎样扩大搜索半径，减少同簇样本数，都只能聚成一类或效果很差，说明此数据集不适宜采用这种聚类方法。

**2.5.3 层次聚类**

* 1. **对比分析和优化**

**2.6.1 模型结果对比分析**

最终各聚类模型评价指标对比如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 模型\指标 | 轮廓系数 | CH Score | DB Score |
| K-Means | 0.2009 | 151.1759 | 1.4152 |
| DBSCAN | 0.0237 | 0.4402 | 55.8094 |
| 层次聚类 |  |  |  |
| GMM | 0.2598 | 171.9434 | 1.5857 |

* 各模型的结果对比分析如下：

K-Means的轮廓系数为0.2009，CH 指数为151.1759，DB 指数为1.4152。轮廓系数较低，表明簇内的紧凑性和簇间的分离度一般，可能是由于数据本身的分布特性并不完全符合K-Means假设的球形簇。然而K-Means在CH 指数上表现较好，说明簇间的分离度和簇内的紧凑性在一定程度上是合理的。DB 指数也相对较低，进一步支持了这一点。总体来看，K-Means在当前数据集上表现中等，能够较好地分离簇，但在簇内紧凑性方面有待提高。

DBSCAN的轮廓系数仅为0.0237，CH 指数为0.4402，DB 指数为55.8094。这些结果表明DBSCAN在这次聚类任务中的表现较差。轮廓系数接近0，说明簇内的点分布较为松散，簇间的边界模糊不清。CH 指数极低，进一步验证了簇间的分离度和簇内的紧凑性都很差。DB 指数较高，再次表明簇间的差异较小.数据本身的分布可能不具备明显的密度差异，使得DBSCAN难以有效地区分不同的簇。

GMM的轮廓系数为0.2598，CH 指数为171.9434，DB 指数为1.5857。GMM在这三个指标上的表现都优于K-Means和DBSCAN。轮廓系数虽然仍然较低，但相比K-Means有所提升，表明GMM在簇内紧凑性和簇间分离度方面表现更好。CH 指数较高，进一步证实了簇间的分离度和簇内的紧凑性较好。DB 指数也相对较低，说明簇间的差异较大。GMM能够捕捉到数据中的多重模态，适应不同形状和大小的簇，因此在当前数据集上表现最佳。

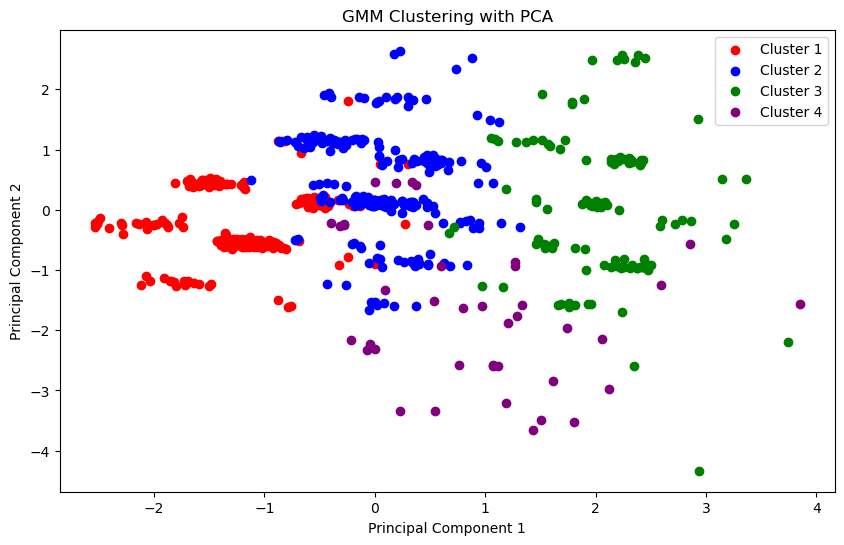
层次聚类

**2.6.2 DBSCAN优化（换用GMM模型）**

由于DBSCAN不适用于本数据集，故考虑换用模型，这里采用高斯混合模型（GMM）聚类。GMM是一种概率模型，它假设数据由多个高斯分布混合而成的，每个高斯分布代表一个潜在的数据类别或簇。GMM不仅能够估计每个数据点属于各个簇的概率，还能提供关于数据分布的更丰富的信息，如每个簇的均值向量和协方差矩阵等。它具有以下特征：

* 软聚类：与K-means等硬聚类算法不同，GMM允许一个数据点同时属于多个簇，每个簇的隶属度由该点在相应高斯分布下的概率密度决定。
* GMM是概率模型，利用最大似然估计确定模型参数，即寻找使数据集出现概率最大的参数值。
* 每个高斯成分有自己的协方差矩阵，因此可适应不同形状和大小的簇。
* GMM可以有效地捕捉数据中的多重模态，即数据可能来自多个不同分布

聚类结果降维后可视化如下：

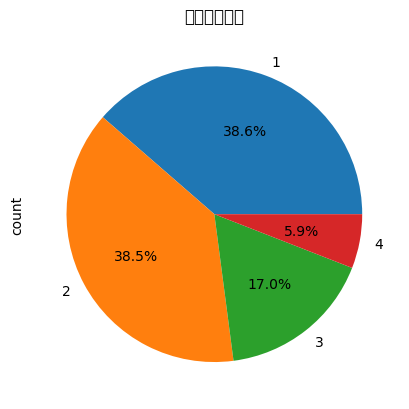


* 相应评价指标：

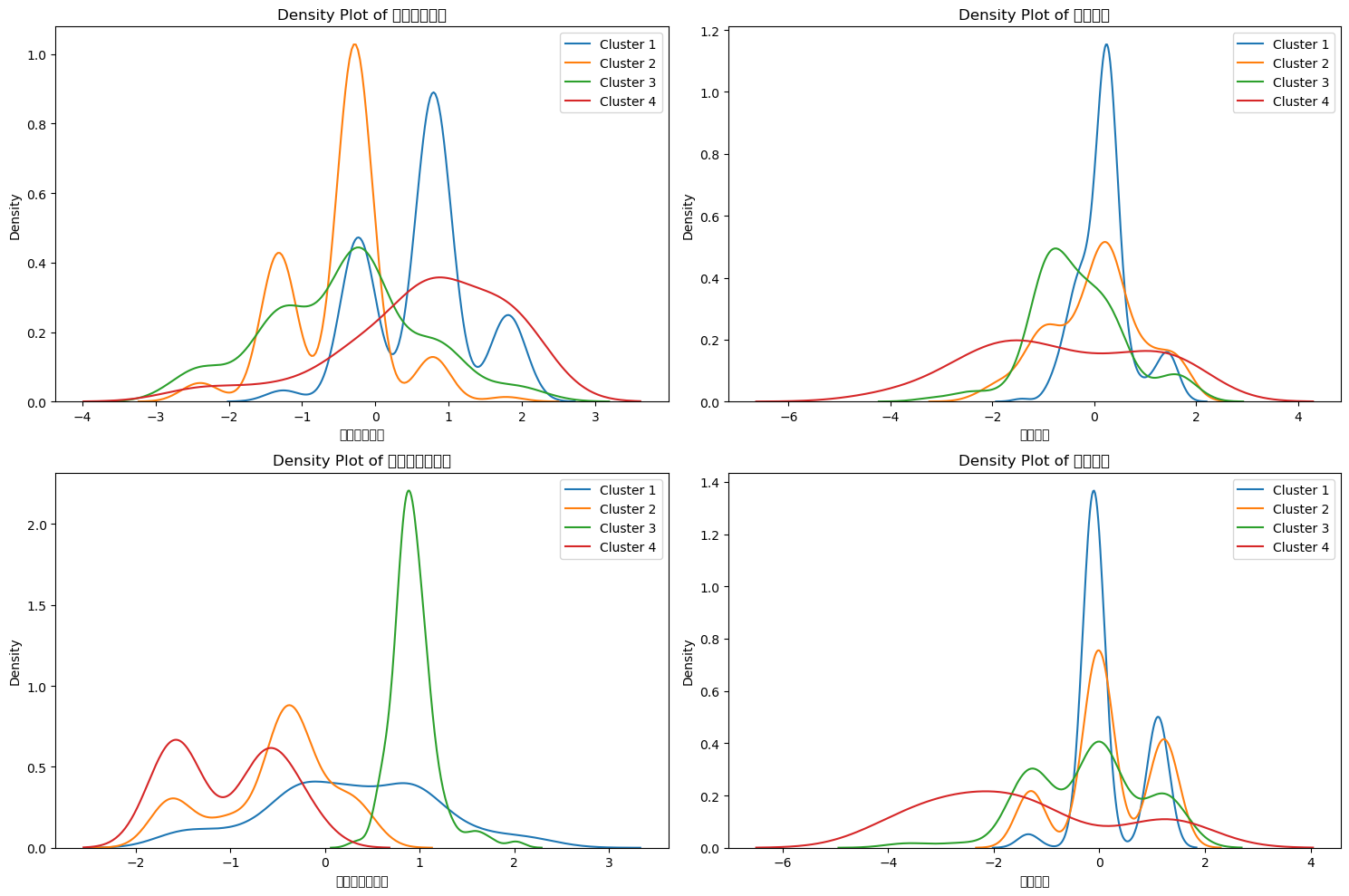
|  |  |  |
| --- | --- | --- |
| Silhouette Coefficient | Calinski-Harabasz Score | Davies-Bouldin Score |
| 0.2598 | 171.9434 | 1.5857 |

尽管轮廓系数较低，但CH指数和DB指数都表明聚类结果在簇间分离和簇内紧凑性方面表现良好，聚类效果较好。这可能是因为：数据本身的分布可能具有明显的多模态特性，适合用GMM进行建模，且选择的参数较为合理。

* 每簇人数占比与结果展示：

* Factor1~4（健康生活习惯,注重品质,创新与消费倾向,规律作息）密度图：



由图中峰值和分布知，GMM聚类结果中健康生活习惯、注重品质对第一类人群影响最显著；健康生活习惯对第二类人群影响最显著；创新与消费倾向对第三类人群影响最显著；健康生活习惯对第四类人群影响最显著。

分布的重叠部分表明类簇在这个变量上的区分度不是很高，而变量上明确的分界线显示出良好的区分能力

* 解释聚类结果：

|  |  |
| --- | --- |
| **用户类型** | **用户特征** |
| **类型一** | 生活习惯健康、注重品质的用户，尝试新事物和新产品倾向一般 |
| **类型二** | 注重健康生活习惯的用户，较注重品质,有一定创新与消费倾向，作息不规律 |
| **类型三** | 对创新和消费的兴趣极高的用户 |
| **类型四** | 生活作息不规律，但有健康生活习惯，创新消费的兴趣一般 |

**2.6.3 层次聚类优化**

**三、实验总结**

1. **总结**