**基于聚类分析的顾客人格分析**

南京大学 信息管理学院 信息管理与信息系统专业

211820117 戴一诺

201830165 倪隽

**摘要**：

**关键词**：聚类分析；人格分析

**1 聚类分析法简介**

聚类分析是一种建立分类的多元统计分析方法，它能够将一批观测（或变量）数据根据其诸多特征，按照在性质上的亲疏程度在没有先验知识的情况下自动分类，产生多个分类结果。类内部个体特征具有相似性，不同类间个体特征的差异性较大。聚类分析通常将每个观测数据视作p维（p个聚类变量）空间上的一个点，点与点之间的距离越小，意味着它们越“亲密”，越有可能聚成一类；点与点之间的距离越大，意味着它们越“疏远”，越有可能分属于不同的类。聚类分析在统计分析的各应用领域得到了广泛应用。

聚类分析主要有层次聚类（Hierarchical Clustering）和K-Means聚类（K-Means Clustering）两种。层次聚类，亦称系统聚类，是指在不同层次对数据集进行划分，从而形成树形的聚类结构。数据集的划分可采用“自底向上”的聚合策略，也可采用“自顶向下”的分拆策略。K-Means聚类，也称快速聚类，是一个反复迭代的分类过程。在聚类过程中，观测所属的类会不断调整，直至最终达到稳定为止。

**2 顾客人格分析数据集简介**

**2.1 顾客人格分析简介**

顾客人格分析是商业分析的重要组成部分。企业常常通过顾客人格分析对企业理想顾客进行详细分析，促进企业对其客户的了解，并帮助企业根据不同类型客户的特定需求、行为和关注点来开发并修正产品。顾客人格分析可以通过帮助企业制定更具针对性的市场推广策略等来节省企业资源、提高企业销售效率。

**2.2 顾客人格分析数据集（Customer Personality Analysis Dataset）简介**

本研究所选用的数据集是kaggle网站上由加利福尼亚大学伯克利分校哈斯商学院（Haas School of Business, University of California Berkeley）教授奥马尔·罗梅罗-埃尔南德斯博士（Dr. Omar Romero-Hernandez）提供的开源数据集：顾客人格分析数据集（Customer Personality Analysis Dataset）。

这个数据集提供的顾客信息主要包括顾客属性、产品购买情况、促销活动参与情况和购买方式四方面。顾客属性主要包括顾客ID（ID）、出生年份（Year of Birth）、教育水平（Education）、婚姻状况（Marital Status）、年收入（Income）以及家庭中儿童和青少年的数量（Kidhome/Teenhome）等；产品购买情况主要包括顾客在过去两年中购买葡萄酒、水果、肉类、鱼类、糖果和黄金等产品的金额；促销活动参与情况主要包括顾客在多个促销活动中是否接受了优惠、是否投诉以及最后一次购买的反应等；购买方式主要包括顾客通过公司网站（Web）、目录（Catalog）和实体店（Store）购买的频次以及在最近一个月内访问公司网站的频率等。

**3 顾客人格聚类分析**

**3.1 研究方法**

**3.1.1 聚类算法**

聚类(Clustering)算法是数据挖掘领域的重要技术，可以描述大量数据的分布情况，发现数据中隐藏的规律。往往可以在处理庞大数据时运用聚类方法把数据分成若干组，并且对分组分析。常见聚类算法包括：分割式聚类算法、系统聚类算法、模糊聚类算法、K-Means聚类算法等。

**3.1.2 K-Means聚类分析**

K-Means(MacQueen,1967)聚类是一种聚类算法，又名为“动态聚类方法”、“逐步聚类方法”，为无监督学习算法之一。K-Means聚类算法简单、易于描述，处理大规模数据集时效率较高，在客户细分领域有着广泛的应用与快速的发展。包括电信(汪亚鸥,2004)、金融(焦瑾,2003)等领域的客户细分业务分析，国内学者都有所实践涉猎[[1]](#footnote-0)。

K-Means聚类的基本步骤如下：

STEP1：确定K值以及初始化聚类中心，选择K个初始凝聚点为计划形成类的中心;

STEP2：计算每一个观测到K个初始凝聚点的距离，将每个观测和最近的凝聚点分到一组，形成K个初始分类;

STEP3：计算初始分类的重心（或均值）为新的凝聚点，重新计算每一个观测到初始分类重心的距离，将每个观测和最近的凝聚点分为一组;

STEP4：重复进行STEP2和STEP3，直至初始分类的重心（或均值）没有明显变化为止。

本次实证分析的研究流程如下(fig1)：

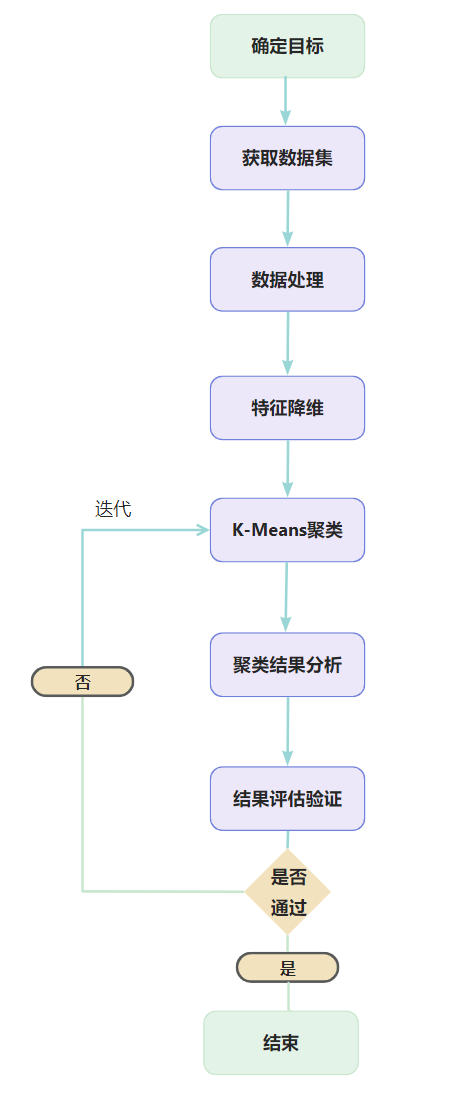


fig1

**3.1.3 主成分分析**

主成分分析(PCA)是对高维数据进行降维的一种技术，利用原始变量的线性组合生成新变量（即主成分），既保留原始变量的大部分信息又相互独立，可以提高聚类的效果。

客户细分的变量一般在几十个到上百个不等，这些变量之间存在较强的相关性，有信息冗余，需要降维；降低维度也能提高聚类的效果。因此对于高维数据，在聚类前需要降低维度，减少参与聚类的变量个数。

**3.2 数据准备**

**3.2.1 数据描述**

所选取数据集来自加利福尼亚大学伯克利分校哈斯商学院（Haas School of Business, University of California Berkeley）教授奥马尔·罗梅罗-埃尔南德斯博士（Dr. Omar Romero-Hernandez）的开源数据集：顾客人格分析数据集（Customer Personality Analysis Dataset）。

数据集属性如下：

|  |  |  |
| --- | --- | --- |
| **客户属性部分** | | |
| **字段名【英】** | **字段名【中】** | **含义** |
| ID | 编号 | Customer's unique identifier |
| Year\_Birth | 出生年份 | Customer's birth year |
| Education | 受教育程度 | Customer's education level |
| Marital\_Status | 婚姻状况 | Customer's marital status |
| Income | 收入 | Customer's yearly household income |
| Kidhome | 家中儿童数量 | Number of children in customer's household |
| Teenhome | 家中青少年数量 | Number of teenagers in customer's household |
| Dt\_Customer | 注册时间 | Date of customer's enrollment with the company |
| Recency | 最近购物时间 | Number of days since customer's last purchase |
| Complain | 投诉 | 1 if the customer complained in the last 2 years, 0 otherwise |

|  |  |  |
| --- | --- | --- |
| **产品属性部分** | | |
| **字段名【英】** | **字段名【中】** | **含义** |
| NumDealsPurchases | 打折购买次数 | Number of purchases made with a discount |
| AcceptedCmp1 | 第一次接受 | 1 if customer accepted the offer in the 1st campaign, 0 otherwise |
| AcceptedCmp2 | 第二次接受 | 1 if customer accepted the offer in the 2nd campaign, 0 otherwise |
| AcceptedCmp3 | 第三次接受 | 1 if customer accepted the offer in the 3rd campaign, 0 otherwise |
| AcceptedCmp4 | 第四次接受 | 1 if customer accepted the offer in the 4th campaign, 0 otherwise |
| AcceptedCmp5 | 第五次接受 | 1 if customer accepted the offer in the 5th campaign, 0 otherwise |
| Response | 最后一次接受 | 1 if customer accepted the offer in the last campaign, 0 otherwise |

|  |  |  |
| --- | --- | --- |
| **购买地属性部分** | | |
| **字段名【英】** | **字段名【中】** | **含义** |
| NumWebPurchases | 公司网站-购买次数 | Number of purchases made through the company’s website |
| NumCatalogPurchases | 购物目录-购买次数 | Number of purchases made using a catalogue |
| NumStorePurchases | 商店-购买次数 | Number of purchases made directly in stores |
| NumWebVisitsMonth | 上月访问公司网站次数 | Number of visits to company’s website in the last month |

共有2240条数据（数据集局部）：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **0** | **1** | **2** | **3** | **4** |
| **ID** | 5524 | 2174 | 4141 | 6182 | 5324 |
| **Year\_Birth** | 1957 | 1954 | 1965 | 1984 | 1981 |
| **Education** | Graduation | Graduation | Graduation | Graduation | PhD |
| **Marital\_Status** | Single | Single | Together | Together | Married |
| **Income** | 58138 | 46344 | 71613 | 26646 | 58293 |
| **Kidhome** | 0 | 1 | 0 | 1 | 1 |
| **Teenhome** | 0 | 1 | 0 | 0 | 0 |
| **Dt\_Customer** | 04-09-2012 | 08-03-2014 | 21-08-2013 | 10-02-2014 | 19-01-2014 |
| **Recency** | 58 | 38 | 26 | 26 | 94 |
| **MntWines** | 635 | 11 | 426 | 11 | 173 |
| **...** | ... | ... | ... | ... | ... |
| **NumWebVisitsMonth** | 7 | 5 | 4 | 6 | 5 |
| **AcceptedCmp3** | 0 | 0 | 0 | 0 | 0 |
| **AcceptedCmp4** | 0 | 0 | 0 | 0 | 0 |
| **AcceptedCmp5** | 0 | 0 | 0 | 0 | 0 |
| **AcceptedCmp1** | 0 | 0 | 0 | 0 | 0 |
| **AcceptedCmp2** | 0 | 0 | 0 | 0 | 0 |
| **Complain** | 0 | 0 | 0 | 0 | 0 |
| **Z\_CostContact** | 3 | 3 | 3 | 3 | 3 |
| **Z\_Revenue** | 11 | 11 | 11 | 11 | 11 |
| **Response** | 1 | 0 | 0 | 0 | 0 |

5 rows × 29 columns

统计数据类型与非空行：

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **字段** | **非空行计数** | **数据类型** |
| 0 | ID | 2240 | int64 |
| 1 | Year\_Birth | 2240 | int64 |
| 2 | Education | 2240 | object |
| 3 | Marital\_Status | 2240 | object |
| 4 | Income | 2216 | float64 |
| 5 | Kidhome | 2240 | int64 |
| 6 | Teenhome | 2240 | int64 |
| 7 | Dt\_Customer | 2240 | object |
| 8 | Recency | 2240 | int64 |
| 9 | MntWines | 2240 | int64 |
| 10 | MntFruits | 2240 | int64 |
| 11 | MntMeatProducts | 2240 | int64 |
| 12 | MntFishProducts | 2240 | int64 |
| 13 | MntSweetProducts | 2240 | int64 |
| 14 | MntGoldProds | 2240 | int64 |
| 15 | NumDealsPurchases | 2240 | int64 |
| 16 | NumWebPurchases | 2240 | int64 |
| 17 | NumCatalogPurchases | 2240 | int64 |
| 18 | NumStorePurchases | 2240 | int64 |
| 19 | NumWebVisitsMonth | 2240 | int64 |
| 20 | AcceptedCmp3 | 2240 | int64 |
| 21 | AcceptedCmp4 | 2240 | int64 |
| 22 | AcceptedCmp5 | 2240 | int64 |
| 23 | AcceptedCmp1 | 2240 | int64 |
| 24 | AcceptedCmp2 | 2240 | int64 |
| 25 | Complain | 2240 | int64 |
| 26 | Z\_CostContact | 2240 | int64 |
| 27 | Z\_Revenue | 2240 | int64 |
| 28 | Response | 2240 | int64 |

数据类型统计： float64(1), int64(25), object(3)

**3.2.2 数据清洗**

删除数据集中的缺失值(NA)：



JupyterNotebook 代码运行界面

删除缺失值后，数据量为2216条。

确定用户注册时间范围：

首先将"Dt\_Customer"列中的日期字符串转换为Python中的日期时间对象，然后找出用户注册最早/晚时间。



JupyterNotebook 代码运行界面

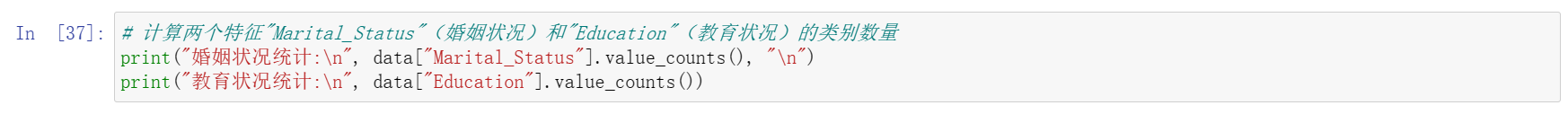
得以确定记载用户最晚注册时间: 2014-06-29；记载用户最早注册时间: 2012-07-30。

记录日期间隔数据（客户注册时长）——创建一个新特征值 "Customer\_For"：



JupyterNotebook 代码运行界面

统计客户婚姻状况与受教育情况：



JupyterNotebook 代码运行界面

统计得到

|  |  |
| --- | --- |
| **婚姻状况统计** | |
| Married | 857 |
| Together | 573 |
| Single | 471 |
| Divorced | 232 |
| Widow | 76 |
| Alone | 3 |
| Absurd | 2 |
| YOLO | 2 |
| 字段名：Marital\_Status, 数据类型：int64 | |

|  |  |
| --- | --- |
| **教育状况统计** | |
| Graduation | 1116 |
| PhD | 481 |
| Master | 365 |
| 2nCycle | 200 |
| Basic | 54 |
| 字段名：Education, 数据类型：int64 | |

**3.2.3 特征提取**

客户年龄计算——新建字段“Age”：



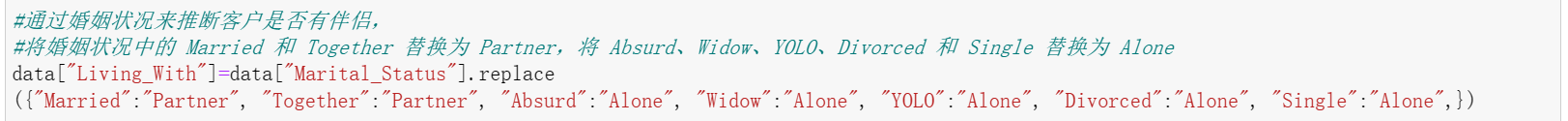
客户各个消费类别的花费总和计算——新建字段“Spent”：

将酒类、蔬果、渔产品等花费求和可得新字段“Spent”。



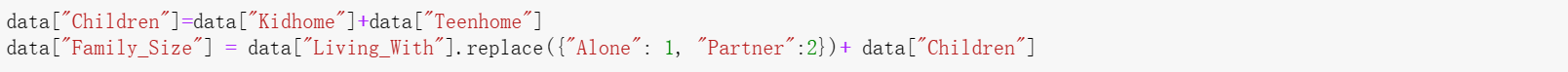
通过婚姻状况来推断客户是否有伴侣——新建字段“Living\_with”：

将婚姻状况中的 Married 和 Together 替换为 Partner，将 Absurd、Widow、YOLO、Divorced 和 Single 替换为 Alone。



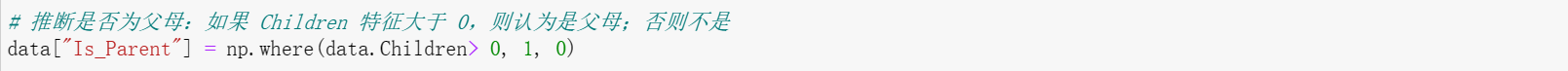
计算客户家庭成员人数——新建字段“Family\_Size”：

上文将 Living\_With 特征中的 Partner 认为是两人，将 Alone 认为是一人，接下来将 Children 特征（Kidhome 和 Teenhome 求和）作为孩子数，最终将 Living\_With 和 Children 求和得到 Family\_Size 特征。



推断客户是否为父母：

如果 Children 特征大于 0，则认为是父母；否则，认为不是。

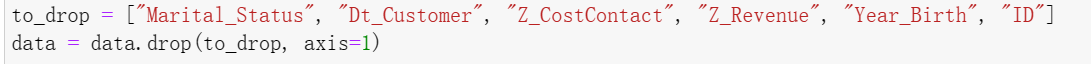


为减轻视觉负担，将重复冗长的字段进行重命名：

把各个消费类别的名称改为简单的名字，例如"MntWines":改为"Wines"，"MntFruits"改为"Fruits"等。



删除一些对于聚类分析无用的特征所在列：Marital\_Status、Dt\_Customer、Z\_CostContact、Z\_Revenue、Year\_Birth 和 ID。



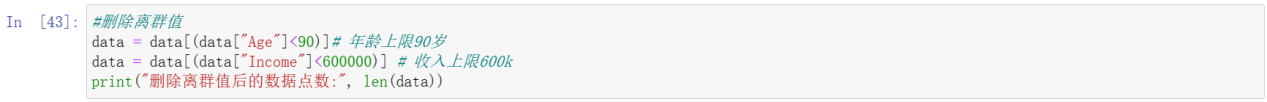
经过特征提取之后，新数据集如下（数据集局部）：

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 2216 | 2216 | 2216 | 2216 | 2216 | 2216 | 2216 | 2216 | 2216 | 2216 | ... | 2216 | 2216 | 2216 | 2216 | 2.22E+03 | 2216 | 2216 | 2216 | 2216 | 2216 |
| **mean** | 52247.25135 | 0.441787 | 0.505415 | 49.012635 | 305.091606 | 26.356047 | 166.995939 | 37.637635 | 27.028881 | 43.965253 | ... | 0.064079 | 0.013538 | 0.009477 | 0.150271 | 3.05E+16 | 52.179603 | 607.075361 | 0.947202 | 2.592509 | 0.71435 |
| **std** | 25173.07666 | 0.536896 | 0.544181 | 28.948352 | 337.32792 | 39.793917 | 224.283273 | 54.752082 | 41.072046 | 51.815414 | ... | 0.24495 | 0.115588 | 0.096907 | 0.357417 | 1.75E+16 | 11.985554 | 602.900476 | 0.749062 | 0.905722 | 0.451825 |
| **min** | 1730 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0.00E+00 | 25 | 5 | 0 | 1 | 0 |
| **25%** | 35303 | 0 | 0 | 24 | 24 | 2 | 16 | 3 | 1 | 9 | ... | 0 | 0 | 0 | 0 | 1.56E+16 | 44 | 69 | 0 | 2 | 0 |
| **50%** | 51381.5 | 0 | 0 | 49 | 174.5 | 8 | 68 | 12 | 8 | 24.5 | ... | 0 | 0 | 0 | 0 | 3.07E+16 | 51 | 396.5 | 1 | 3 | 1 |
| **75%** | 68522 | 1 | 1 | 74 | 505 | 33 | 232.25 | 50 | 33 | 56 | ... | 0 | 0 | 0 | 0 | 4.57E+16 | 62 | 1048 | 1 | 3 | 1 |
| **max** | 666666 | 2 | 2 | 99 | 1493 | 199 | 1725 | 259 | 262 | 321 | ... | 1 | 1 | 1 | 1 | 6.04E+16 | 128 | 2525 | 3 | 5 | 1 |
|  | **Income** | **Kidhome** | **Teenhome** | **Recency** | **Wines** | **Fruits** | **Meat** | **Fish** | **Sweets** | **Gold** | **...** | **AcceptedCmp1** | **AcceptedCmp2** | **Complain** | **Response** | **Customer\_For** | **Age** | **Spent** | **Children** | **Family\_Size** | **Is\_Parent** |

8 rows × 28 columns

**3.2.4 数据标准化**

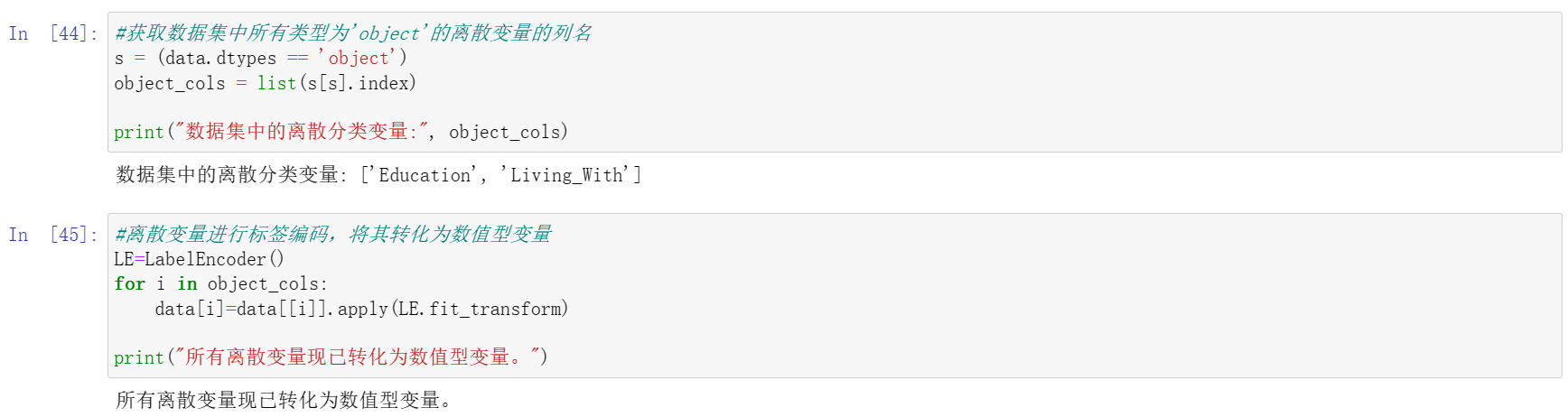
删除离群值：



删除离群值后的数据点数: 2212

将所有离散变量转化为数值型变量：

先找出离散变量，再进行标签编码。



将数据整体缩放：



**3.3 特征降维**

利用主成分分析(PCA)将28个属性进行降维，根据尝试与经验发现降至3维不仅不损失太多关键信息，也更方便后续三维空间中的图表观察。

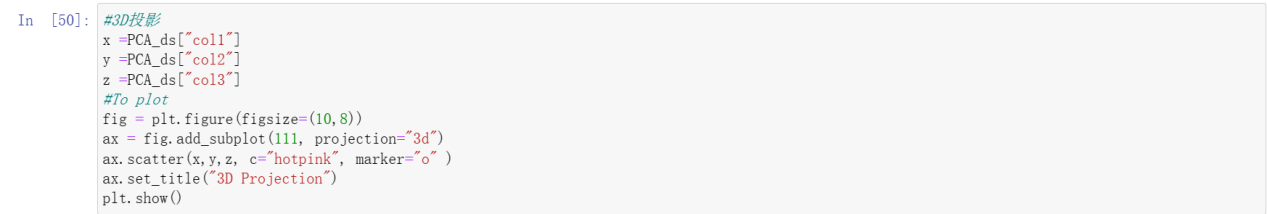


得到降维后的数据：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **col1** | 2212 | 6.99E-17 | 2.878602 | -5.978123 | -2.53947 | -0.781595 | 2.38638 | 7.452915 |
| **col2** | 2212 | 2.51E-17 | 1.709469 | -4.194757 | -1.323932 | -0.173716 | 1.234923 | 6.168185 |
| **col3** | 2212 | 7.21E-17 | 1.231685 | -3.625184 | -0.853556 | -0.051292 | 0.863841 | 6.746845 |

在3D投影图中观察数据分布(fig2)：

在三维空间中观察不同的数据点和它们所属的聚类，从而更好地理解数据的聚类效果。



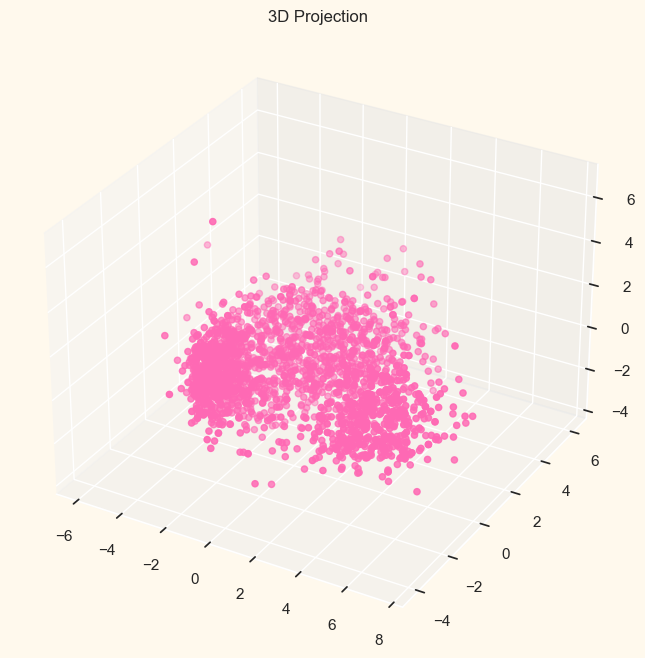
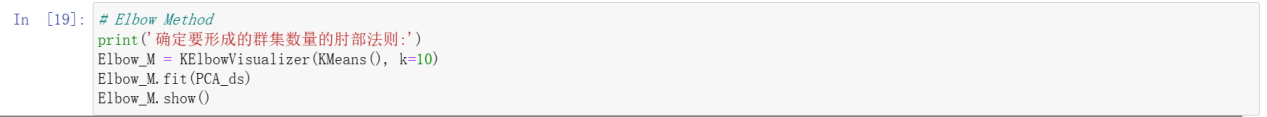


fig3

**3.4 K-Means聚类**

使用KElbowVisualizer可视化来确定最佳的聚类数目：

采用“肘部法则”(Elbow Method)，对不同聚类数下的K-Means模型拟合结果进行评估，绘制出损失函数随着聚类数的增加而下降的趋势，找到一个“拐点”（即损失函数曲线下降速度明显变缓的位置），认为这个位置对应的聚类数是最优的。这个“拐点”就像是手肘一样，因此被称为“肘部”。



确定要形成的群集数量的肘部法则(fig3):

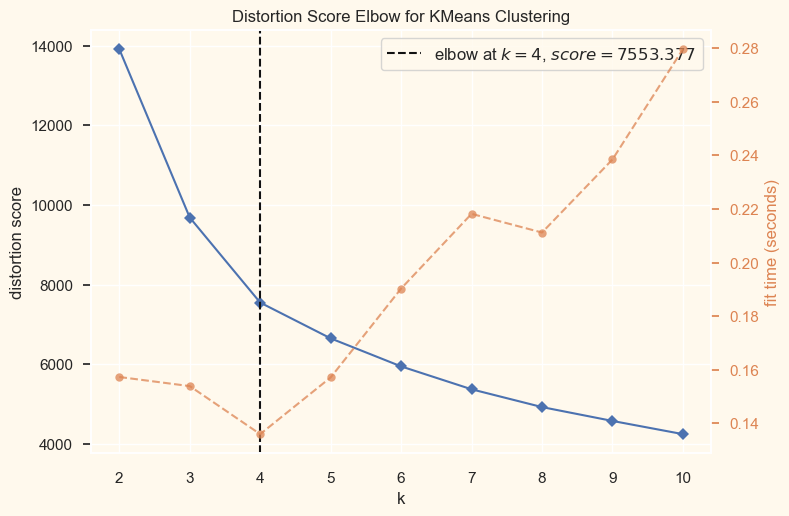
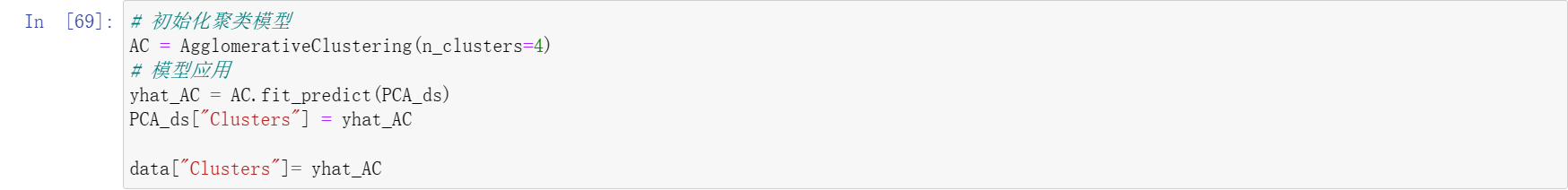


fig3.<AxesSubplot:title={'center':'Distortion Score Elbow for KMeans Clustering'}, xlabel='k', ylabel='distortion score'>

显然，当k=4是“拐点”，即该位置对应的聚类数4是最优的。因此确定本实证研究的聚类数目为4。

使用凝聚层次聚类模型对降维后的数据进行聚类，分为4类，并将聚类的结果作为新特征"Clusters"添加到原始数据集"data"中：

用 AgglomerativeClustering 初始化一个凝聚层次聚类模型，聚类数设为4。然后使用 fit\_predict() 方法将模型应用于降维后的数据 PCA\_ds 上，生成聚类标签 yhat\_AC。然后将聚类标签 yhat\_AC 添加到 PCA\_ds 数据框中的 "Clusters" 列中，同时也添加到原始数据 data 中的 "Clusters" 列中。



**3.5 聚类分析可视化**

将数据在三维上可视化：

将数据点按照聚类结果着色，参数c=PCA\_ds["Clusters"] 表示将聚类结果作为颜色的依据。最后使用matplotlib的3D散点图绘制函数scatter绘制出可视化图形。



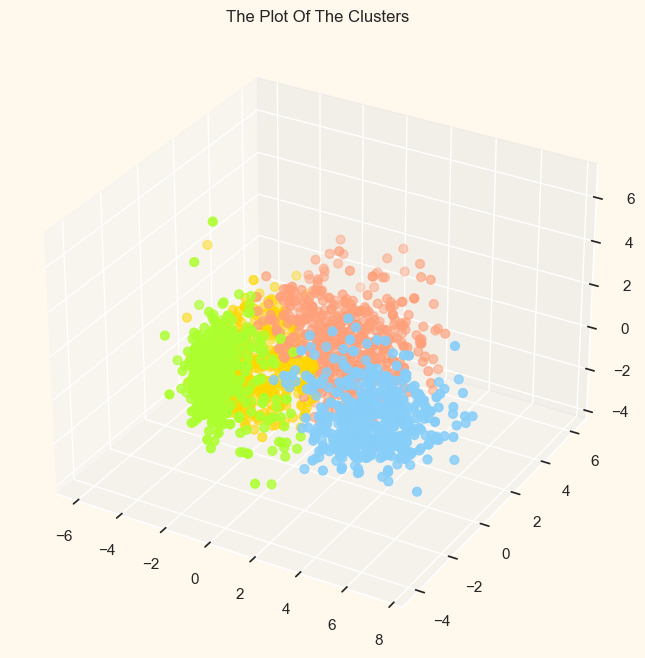


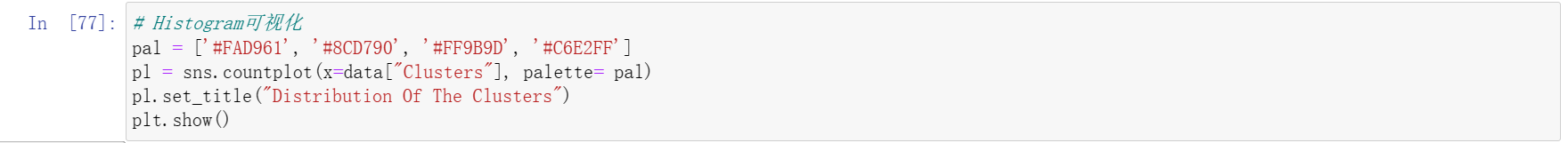
fig4.The Plot Of The Clusters

从图中(fig4)可以看出，数据已经按照四种不同的颜色进行了聚类。

直方图可视化：

将四类数据的直方图绘制出来，分别是对应四种颜色。

可以(fig5)发现四类客户在数据量上大致是均匀分布的、相当的。



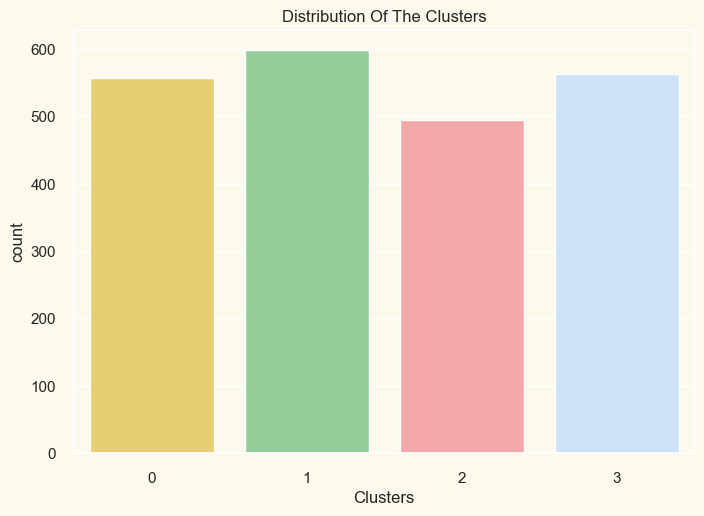
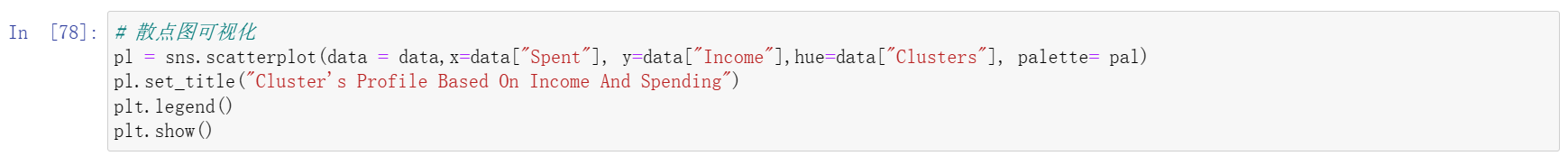


fig5.Histogram

绘制散点图(fig6)：

将经过聚类的数据集按照收入和花费进行可视化展示，每个点的颜色表示该点所属的聚类。帮助我们了解不同聚类的特点和差异，以便更好地理解每个聚类所代表的群体的收支习惯特征。



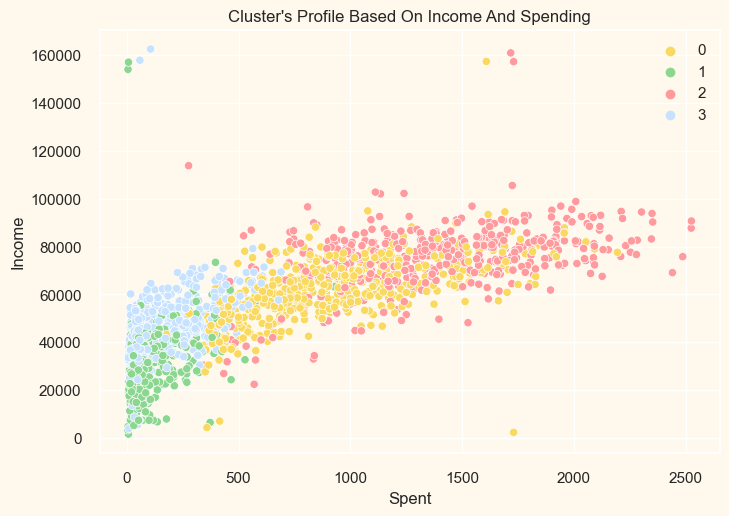


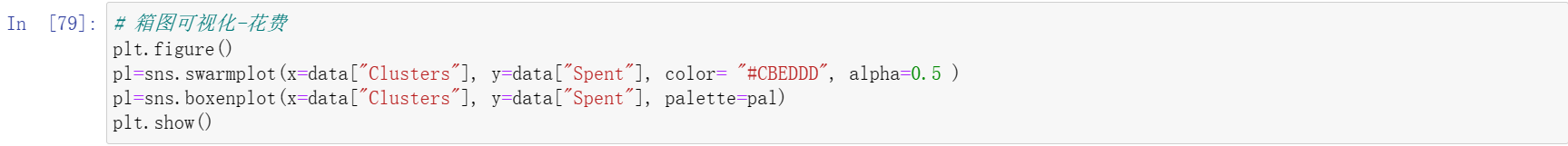
fig6.各类客户收支散点图

绘制箱线图、蜂群图：

将聚类结果可视化为箱线图和蜂群图(fig7)，展示了不同聚类中客户的支出情况。

其中，x轴为聚类结果，y轴为客户的支出。

箱线图展示了不同聚类中客户支出的分布情况，包括中位数、四分位数、离群点等信息；蜂群图则以点的形式展示了每个客户的支出值，以更直观地展示分布情况。两种图都以颜色区分不同的聚类。



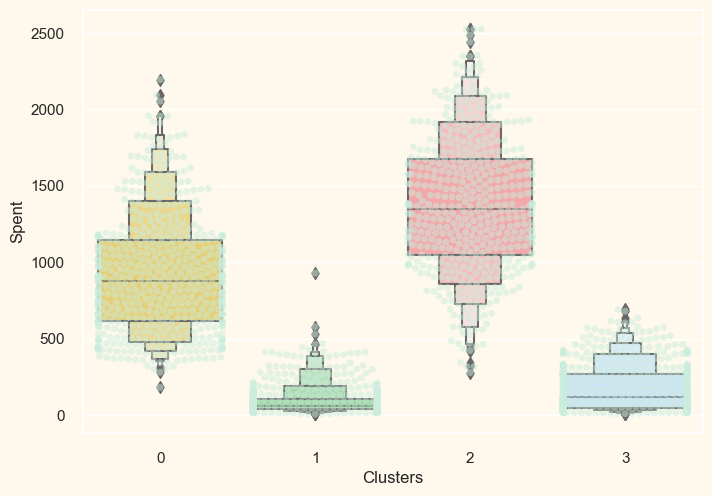


fig7.箱线图和蜂群图-客户支出情况

将聚类结果可视化为箱线图(fig8)，展示了不同聚类中客户的购物情况。



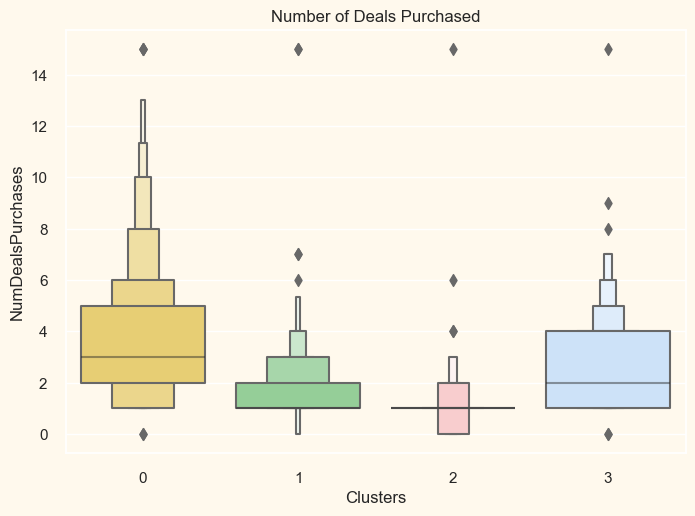


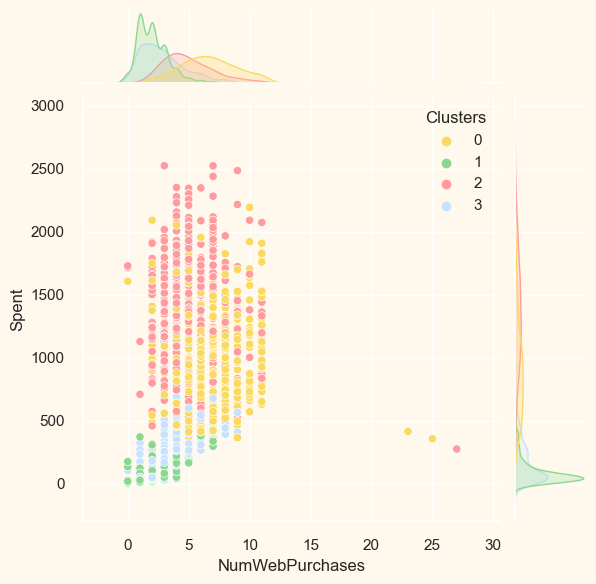
fig8.箱线图-客户打折购物次数

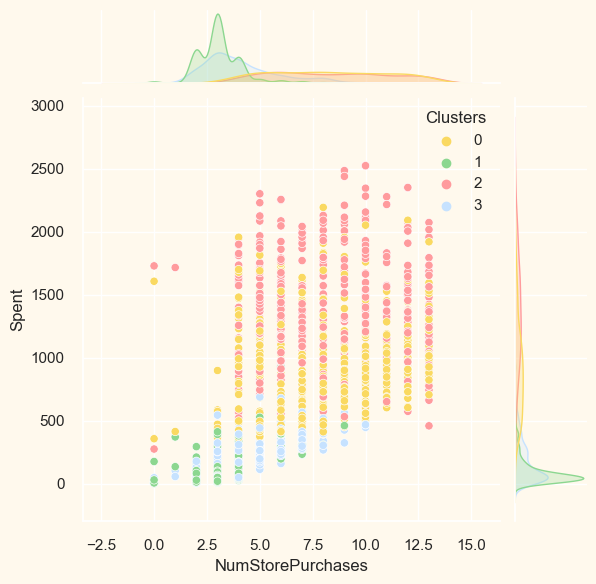
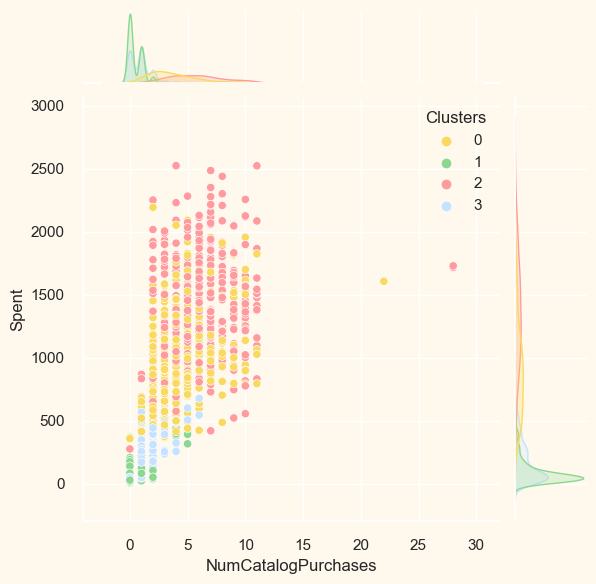
分析不同聚类中顾客的消费方式：

首先指定了四个消费方式变量"NumWebPurchases", "NumCatalogPurchases", "NumStorePurchases", "NumWebVisitsMonth"（公司网站-购买次数、购物目录-购买次数、商店-购买次数、上月访问公司网站次数），然后通过循环将每个消费方式与顾客花费数据进行二元散点图可视化(fig9)，并用不同的颜色表示不同聚类中的顾客。

用于帮助我们了解不同聚类中顾客的消费行为差异，以及哪些消费方式与顾客花费之间存在较强的相关性。







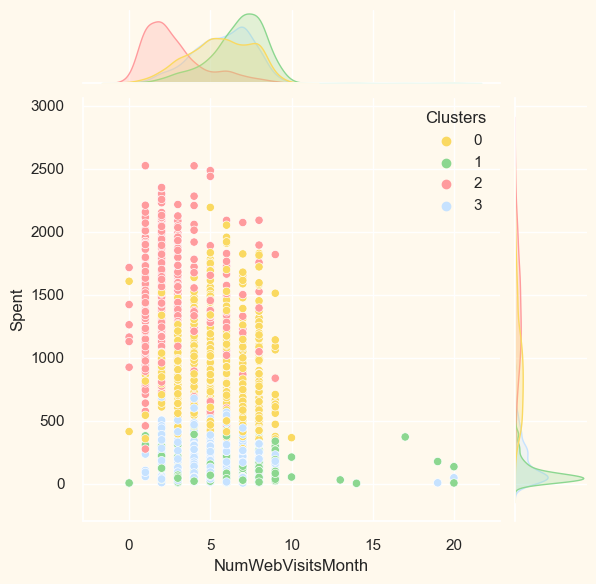
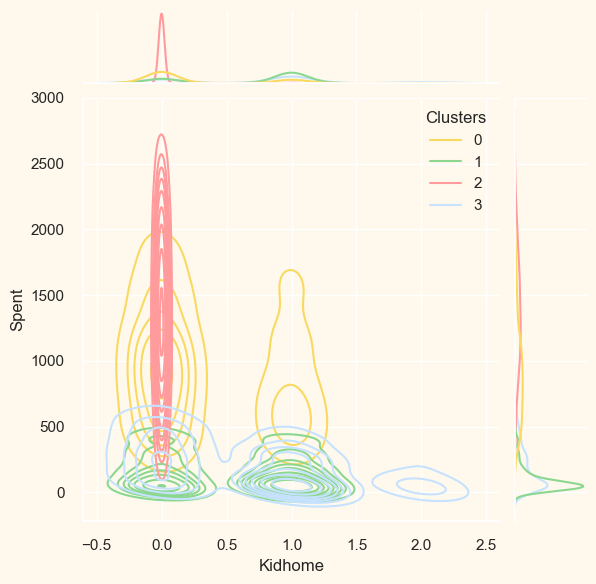


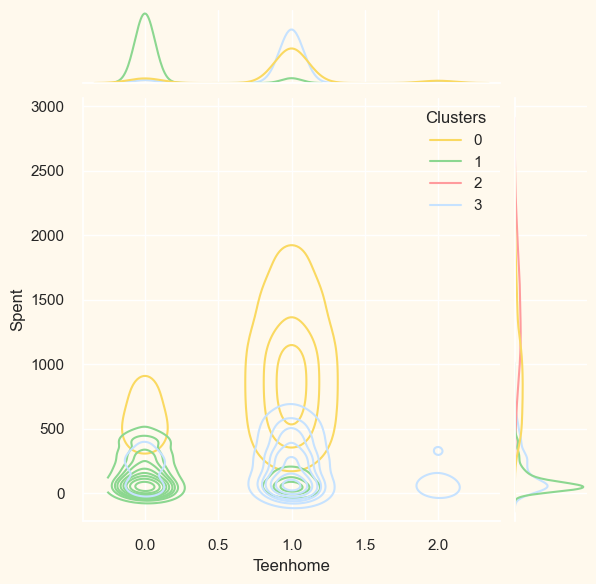
fig9.消费方式分析

客户画像：

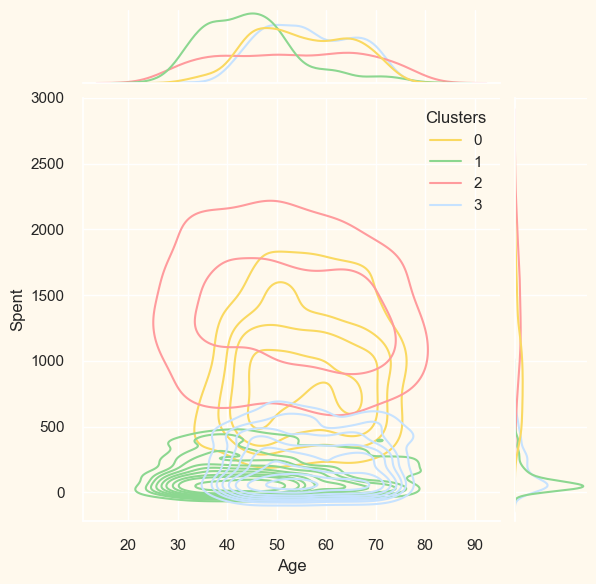
分别通过"Kidhome", "Teenhome", "Customer\_For", "Age", "Children", "Family\_Size", "Is\_Parent", "Education","Living\_With"（家中孩子数量、育儿情况、年龄、家庭大小、教育状况等）等字段可画不同聚类的用户画像(fig10)。

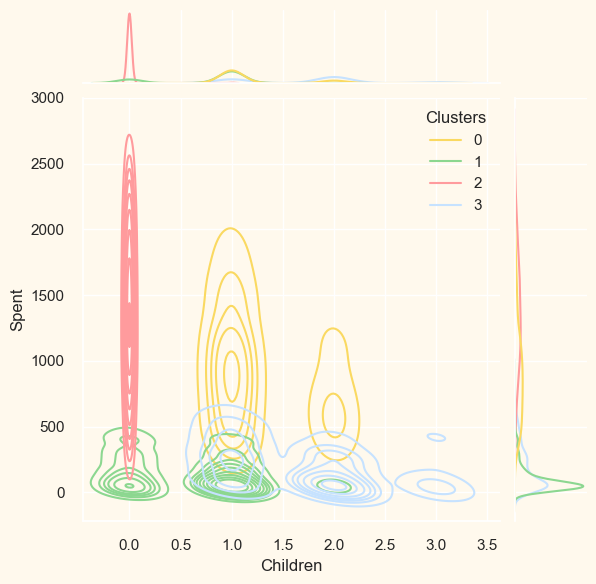


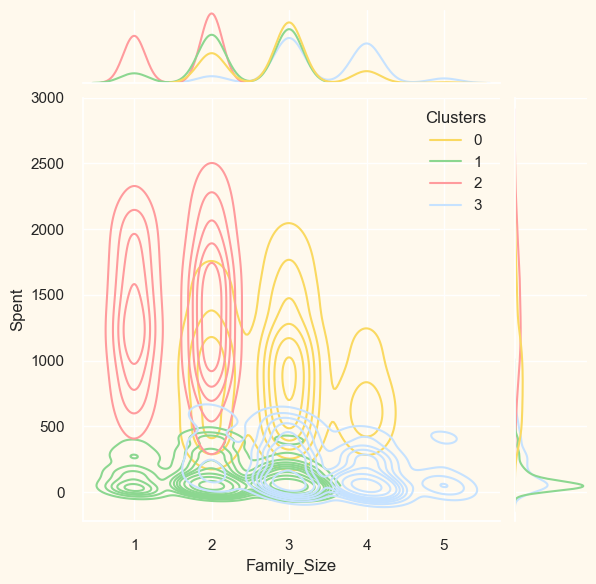


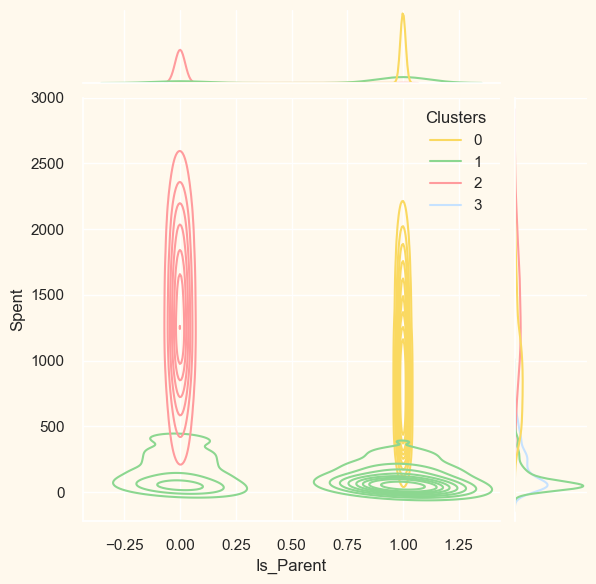


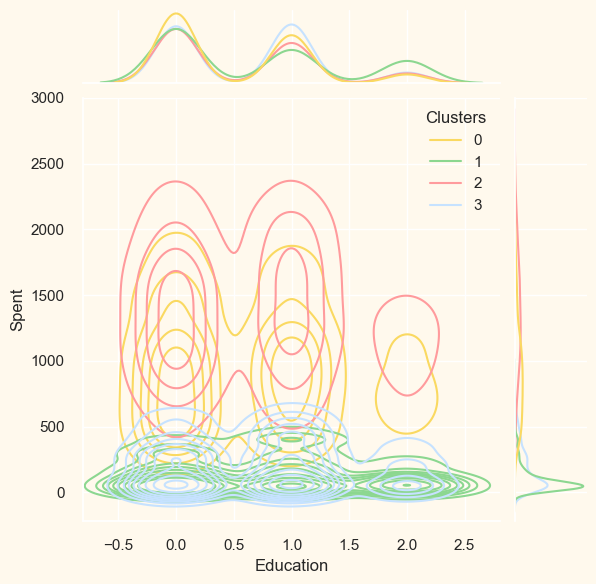












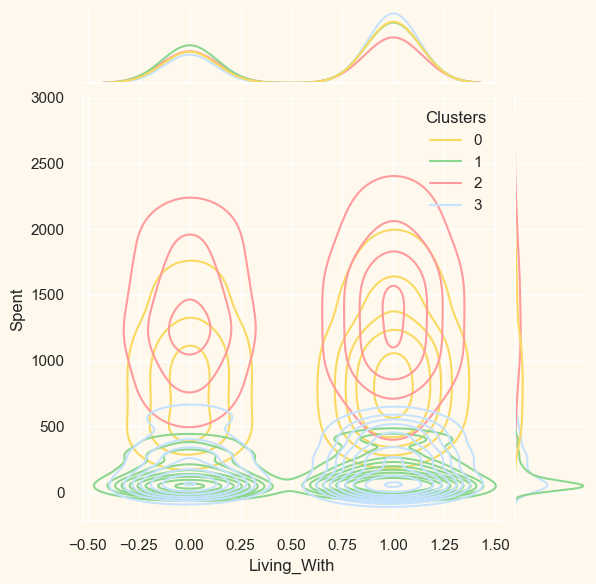


fig10.客户画像

**3.6 顾客人格聚类分析**

根据K-Means聚类，将客户群分为4类，分别命名为0、1、2、3。

首先从整体数量上来看，四类客户群在数据量上大致是均匀分布的、相当的。

其次，从收入-支出情况来看，客户群2的一般收入与支出均位于高位，客户群0次之，再次的客户群3总体比1收入更高、支出稍多，客户群1总体收入与支出最低。

再次，观察不同客户群体的打折购物次数：客户群0打折购物次数最多，其中50%客户的购买次数在2~5次，最多的达到13次；其次客户群3的50%客户的打折活动购买次数在1~4次，以2次为中位数；客户群1的50%客户的活动购买次数在2次左右，客户群2则约为0~2次。

此外，就购物方式而言，客户群0、2更偏爱网络购物和购物目录购物，客户群1、3对于线下商店购物的热情与频次高于其他两种非线下实体店购物方式。然而，相对而言客户群1、3访问官网的次数显然多于其他两类客户，也就是说，在线上，客户群1、3的下单率/点击率低，线上成交率低。

综合以上聚类以及可视化分析结果，可以为四类客户群进行画像：

客户群0——为人父母（包含单亲）；家中有2到4个孩子并且最多有一个青少年；年龄比其他组别要稍大；收入较高，开销亦高，购物需求旺盛；热衷于打折购物，善于精打细算；网络购物和购物目录购物更适合繁忙的他们。

客户群1——非父母角色；家庭构成简单，最多2人；有伴侣的客户比起单身独居的居多；覆盖全年龄段；受教育程度高；收入与支出水平一般；喜爱逛实体商店，线上成交率低。

客户群2——大多数已成为父母的年轻客户群体；家庭形态基本上是核心家庭（伴侣二人以及最多一个年幼孩子）；高收入，支出金额大，购买能力最强；打折购物次数少，很少在意打折促销活动，经济因素不是他们最需要考虑的；偏爱网络购物和购物目录购物，更适合繁忙的他们。

客户群3——为人父母，且多处于5人以下的大家庭中；家中有青少年孩子；收入水平较低（仅高于客户群1）；喜爱打折购物，善于精打细算；购物方式分布均匀且线上成交率低。

根据以上分析可以发现，四个客户群体的购物行为与其人格情况（即收入、家庭状况等个人情况）相符合，可以给出较为合理的解释。也就是说，结合数据分析与数据集背景知识，可以双向地映证该聚类结果是合适的、可以接受的。

**4 结语**

通过本次实证分析，应用K-Means等聚类方法以及可视化分析方法，我们对该客户数据集中的2000多位客户进行了客户细分，并且给出了贴切的用户画像，对客户的人格情况与购物习惯进行了可接受的合理解释，同时证实了本次聚类分析的结果可信。

但是考虑到客户细分的真实世界实用背景，除了我们所使用的指标以及经典RFM模型之外，还要考虑到在日新月异的商业进步以及电子商务进步中，还有其他可待整合发掘的聚类角度[[2]](#footnote-1)，可以作为未来研究方向；此外，K-Means算法本身的改进也有越来越多的研究可以考虑[[3]](#footnote-2)，使得聚类结果更能服务于真实的客户细分场景，可以纳入研究点。

参考文献

1. 王千,王成,冯振元等.K-means聚类算法研究综述[J].电子设计工程,2012,20(07):21-24.DOI:10.14022/j.cnki.dzsjgc.2012.07.034.
2. 邢留伟. K-Means算法在客户细分中的应用研究[D].西南财经大学,2007.
3. 黄亚萍,李垣江.基于K-means算法的电子商务客户细分研究[J].电子设计工程,2017,25(02):63-66+70.DOI:10.14022/j.cnki.dzsjgc.2017.02.015.
4. 杜巍,赵春荣,黄伟建.改进的k-means聚类算法在客户细分中的应用研究[J].河北经贸大学学报,2014,35(01):118-121.
5. 覃茂刚,龙根元,李海云,黄海波,陈万利,陈文.基于K-means聚类层次分析模型的中沙环礁地质环境稳定性定量分析[J/OL].热带海洋学报:1-10[2023-05-05].http://kns.cnki.net/kcms/detail/44.1500.p.20220616.1754.004.html

1. 邢留伟. K-Means算法在客户细分中的应用研究[D].西南财经大学,2007. [↑](#footnote-ref-0)
2. 黄亚萍,李垣江.基于K-means算法的电子商务客户细分研究[J].电子设计工程,2017,25(02):63-66+70.DOI:10.14022/j.cnki.dzsjgc.2017.02.015. [↑](#footnote-ref-1)
3. 杜巍,赵春荣,黄伟建.改进的k-means聚类算法在客户细分中的应用研究[J].河北经贸大学学报,2014,35(01):118-121. [↑](#footnote-ref-2)