项目 3

商超用户价值聚类分析

在商超行业中，顾客关系管理（CRM）扮演着至关重要的角色。通过 CRM 系统，商超可以有效地管理和维护与顾客之间的关系，实现精准化运营并获得最大的转化率。其中，顾客画像和顾客分类是 CRM 的核心环节，它们为商超提供了将顾客群体细分为不同类型的低价值顾客和高价值顾客的能力。因此，在商超行业，借助 CRM 系统，结合顾客画像和顾客分类，商超可以实现精准化经营，提升顾客忠诚度和满意度，从而取得更大的市场份额和竞争优势。本项目将利用商超的顾客数据，结合 RFMPI 模型和 K-Means 聚类算法，对顾客进行分群，比较不同类型顾客的顾客价值，进而制定相应的营销策略。

学习目标

1. 技能目标

（1）能够实现数据的质量评估和预处理。

（2）能够使用可视化图表分析信息。

（3）能够分析数据中的相关性。

（4）能够构建改进的 RFM 模型。

（5）能够确定最佳聚类数。

（6）能够创建和训练 K-Means 聚类模型。

（7）能够分析不同顾客群体的特征。

2. 知识目标

（1）了解商超用户价值聚类分析案例的背景、目标和流程。

（2）掌握数据预处理的方法，对数据进行质量评估和预处理。

（3）掌握可视化分析的方法，对顾客基本信息、消费行为、消费渠道和满意度进行分析。

（4）掌握相关性分析的方法，对数值型属性进行相关性分析和筛选。

（5）掌握 RFM 模型的原理，对 RFM 模型进行改进。

（6）掌握聚类数寻优的方法，寻找最优聚类数，并构建和训练 K-Means 聚类模型。

3. 素养目标

（1）通过商超用户价值分析，强化在数据使用过程中遵循伦理规范的重要性，提升对已有数据资源合理利用和保护的意识。

（2）在商超用户信息预处理、聚类分析和价值评估的过程中，锻炼严谨的数据处理与分析

能力，提升针对不同顾客群体制定精准营销策略的科学素养，深化对数据科学在商超行业应用价值及社会责任方面的认识。

情景描述

信息时代的到来标志着企业营销核心从产品中心向顾客中心的转变，顾客关系管理变得尤为重要。在这一背景下，顾客分类成为商超优化营销资源分配的关键问题之一。通过顾客分类，商超可以精准区分低价值顾客和高价值顾客，为不同价值的顾客群体设计个性化服务方案和营销策略，以最大限度地利用有限资源，实现利润最大化。顾客分类结果的准确性对于商超做出决策至关重要，因为它是优化营销资源分配的基础依据，有助于商超建立更有效的顾客关系管理体系。随着信息技术的不断发展，商超能够更精准地进行顾客分类，实现个性化营销，提升顾客满意度，取得更大的市场竞争优势。

在激烈的市场竞争中，各家商超纷纷推出更具吸引力的营销方式，以吸引更多的顾客。盒马鲜生通过建立会员体系，跟踪顾客购买数据并分析其消费习惯，以向会员提供个性化的推荐和促销活动，增强顾客黏性，同时通过积分兑换和优惠等方法激励顾客持续消费，提高顾客忠诚度。山姆超市则采用批发会员模式吸引商超和个人顾客，建立稳定的顾客群体，设立专业的顾客服务团队，提供个性化服务，从而提高顾客满意度和忠诚度。这些策略和方式有助于盒马鲜生和山姆超市在顾客关系管理上不断提升顾客体验和顾客忠诚度，并实现销售额的增长。

项目分解

某商超面临着顾客流失、竞争力下降和未充分利用商超资源等经营危机。为了应对这些挑战，该商超决定建立一个合理的顾客价值评估模型，通过对顾客进行分类，进一步分析和比较不同顾客群体的价值，制定相应的营销策略并提供个性化服务，以便有针对性地利用有限的资源来满足不同价值顾客的需求。通过这样的举措，商超能够更有效地应对市场竞争，提升顾客忠诚度，并实现业务的可持续增长。

结合商超已积累的大量会员档案信息和购物记录，实现以下目标。这样的做法将帮助商超更好地了解顾客需求，提升顾客满意度和忠诚度，从而实现利润的最大化。

（1）利用商超顾客数据进行分类。

（2）对不同顾客类别进行特征分析，比较不同顾客群体的价值。

（3）为不同价值的顾客类别提供个性化服务，并制定相应的营销策略。

商超用户价值聚类分析的总体流程如图 3-1 所示。

商超用户价值聚类分析主要包括以下 3 个步骤。

（1）获取数据。获取注册日期为 2021 年至 2023 年的商超会员消费数据。

（2）数据探索与可视化。对数据进行探索分析与预处理，包括数据质量评估与预处理、可视化分析、相关性分析等操作。

（3）构建聚类模型并分析结果。筛选与构造建模指标用于建立改进的 RFM 模型 RFMPI，通过肘方法（Elbow Method）确定最佳聚类数量，使用 K-Means 聚类算法进行顾客分群，并对模型结果进行分析，分析不同顾客群体的价值。

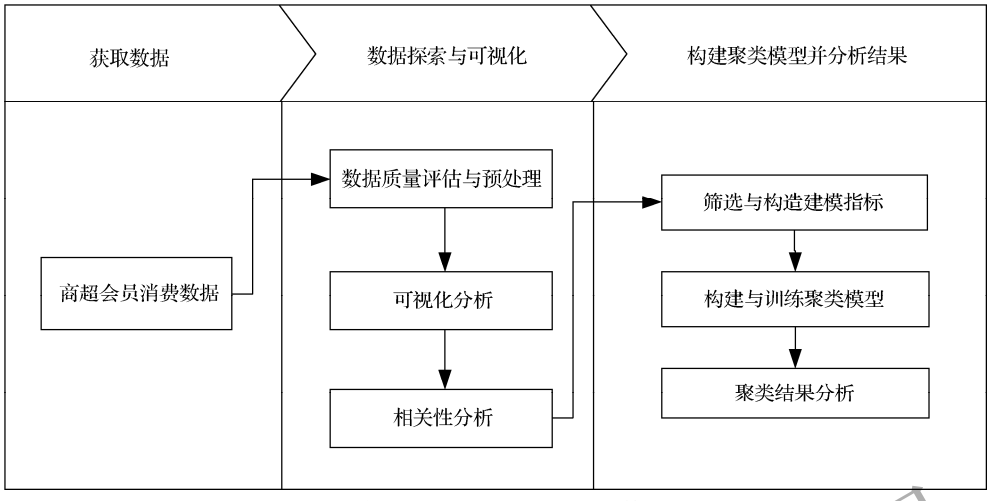

图 3-1　商超用户价值聚类分析总体流程图

项目实施

数据探索与可视化分析

任务 3.1　数据探索与可视化分析

本任务将分为 3 个步骤实现数据探索与可视化分析。首先，对数据进行质量评估与预处理，包括数据描述性分析和数据清洗等步骤，以确保后续分析结果的准确性。其次，对顾客基本信息、消费行为、消费渠道和顾客满意度等因素进行可视化分析，展现数据间的关联和趋势，以理解顾客行为和需求。最后，通过相关性分析揭示不同变量之间的关联程度，以发现潜在的影响因素和规律，从而优化顾客关系管理策略，提升服务质量和顾客满意度。

3.1.1　数据质量评估与预处理

从商超系统的详细数据中，选择宽度为 3 年的时间段作为分析的观测窗口，抽取 2021 年 1 月 1 日至 2023 年 12 月 31 日之间注册成为商超会员的顾客的详细数据。商超会员的记录包括顾客基本信息、注册信息、消费信息、参与促销信息、客诉信息等，共涵盖了 29 个字段，如表 3-1 所示。

表 3-1　商超会员顾客信息字段说明

信息类型	字段名称	说明
顾客基本信息	顾客 ID	顾客的注册 ID
	出生年份	顾客的出生年份
	受教育程度	顾客的受教育程度，如本科、硕士、博士、其他
	婚姻状况	顾客的婚姻状况，如未婚、已婚、离异、丧偶
	年收入/元	顾客的家庭年收入
	儿童数量/人	顾客家庭中的儿童数量
	青少年数量/人	顾客家庭中的青少年数量

续表

信息类型	字段名称	说明
注册信息	注册日期	顾客注册会员的日期
	入会费用/元	入会费用
	注册手续费/元	注册手续费
消费信息	距上次消费天数/天	自顾客上次消费以来的天数
	酒类消费/元	过去 2 年在酒类产品上花费的金额
	水果消费/元	过去 2 年在水果类产品上花费的金额
	肉类消费/元	过去 2 年在肉类产品上花费的金额
	鱼类消费/元	过去 2 年在鱼类产品上花费的金额
	糖果消费/元	过去 2 年在糖果类产品上花费的金额
	黄金消费/元	过去 2 年在黄金产品上花费的金额
	折扣消费次数/次	使用折扣进行消费的次数
	网站消费次数/次	通过公司网站进行消费的次数
	App 消费次数/次	通过公司 App 进行消费的次数
	商店消费次数/次	直接在商店进行消费的次数
	上月网站访问次数/次	上个月访问公司网站的次数
参加促销信息	第 1 次促销	如果顾客参加了第 1 次促销活动，则为 1，否则为 0
	第 2 次促销	如果顾客参加了第 2 次促销活动，则为 1，否则为 0
	第 3 次促销	如果顾客参加了第 3 次促销活动，则为 1，否则为 0
	第 4 次促销	如果顾客参加了第 4 次促销活动，则为 1，否则为 0
	第 5 次促销	如果顾客参加了第 5 次促销活动，则为 1，否则为 0
	最近一次促销	如果顾客参加了最近一次促销活动，则为 1，否则为 0
客诉信息	是否投诉	如果顾客在过去 2 年内有过投诉行为，则为 1，否则为 0

对商超顾客详细数据集进行缺失值评估，查看数据集中的数据缺失情况，并根据缺失情况决定处理方法，如代码 3-1 所示。

代码 3-1　数据缺失值评估

```
import pandas as pd
data = pd.read_excel('../data/商超顾客价值分析数据.xlsx')
# 将注册日期转换为连续型数据
import datetime
today = datetime.date(2023, 12, 31)
data['注册日期'] = data['注册日期'].apply(lambda x:datetime.datetime.strptime(x,
'%Y-%m-%d').date())
# 查看数据缺失情况
# 获取每个字段的缺失值个数
info_missing = data.isnull().sum()
# 获取每个字段的最大值
info_max = data.max()
# 获取每个字段的最小值
info_min = data.min()
# 将结果合并成一个数据框
```

```
info_data = pd.concat([info_missing, info_max, info_min], axis=1)
info_data.columns = ['缺失值个数', '最大值', '最小值']
info_data.describe
```

经检查，发现在 2240 条记录中，仅“年收入/元”字段存在 24 个缺失值，占比 1%，如表 3-2 所示。由于该字段中的缺失值占比较低，因此将删除这些含有缺失的记录。

表 3-2　商超顾客信息描述性分析

字段名称	缺失值个数	最大值	最小值
顾客 ID	0	11191	0
出生年份	0	2009	1906
受教育程度	0	硕士	其他
婚姻状况	0	离异	丧偶
年收入/元	24	162397	1730
儿童数量/人	0	2	0
青少年数量/人	0	2	0
注册日期	0	2023/12/30	2021/1/1
入会费用/元	0	3	3
注册手续费/元	0	11	11
距上次消费天数/天	0	99	0
酒类消费/元	0	1493	0
水果消费/元	0	199	0
肉类消费/元	0	1725	0
鱼类消费/元	0	259	0
糖果消费/元	0	262	0
黄金消费/元	0	321	0
折扣消费次数/次	0	15	0
网站消费次数/次	0	27	0
App 消费次数/次	0	28	0
商店消费次数/次	0	13	0
上月网站访问次数/次	0	20	0
第 1 次促销	0	1	0
第 2 次促销	0	1	0
第 3 次促销	0	1	0
第 4 次促销	0	1	0
第 5 次促销	0	1	0
最近一次促销	0	1	0
是否投诉	0	1	0

此外，为方便后续分析对原始数据进行预处理，包括以下 3 个处理步骤，如代码 3-2 所示。

（1）将顾客出生年份信息转换为年龄信息。

（2）将顾客注册日期信息转换为注册天数信息。

（3）由于入会费用和注册手续费均相同，因此对这两行进行剔除处理。

代码 3-2　数据预处理

```
# 处理缺失值
data = data.dropna()  # 删除包含缺失值的样本
# 数据预处理
data['年龄'] = 2023 - data['出生年份']
data['注册天数/天'] = data['注册日期'].apply(lambda x:(today - x).days)
data.drop(columns=['入会费用/元', '注册手续费/元'], inplace=True)
data.to_csv('../tmp/数据预处理.csv', index=False, encoding='utf-8-sig')
```

3.1.2　可视化分析

可视化分析在深入挖掘顾客基本信息、消费行为、消费渠道及顾客满意度等数据方面扮演着至关重要的角色。可视化分析能够直观展现顾客群体的特征、购买模式、渠道偏好及满意度波动趋势，从而为营销策略的制定和顾客体验的优化提供坚实的数据支撑。

1. 顾客基本信息

顾客基本信息中的年龄、年收入、受教育程度和婚姻状况是商超了解顾客特征和需求的重要指标。这些信息能够帮助商超更好地制定个性化营销策略，满足不同顾客群体的需求，从而提供更具针对性的产品和服务。

1）年龄

绘制柱状图分析商超顾客的年龄情况，如代码 3-3 所示。

代码 3-3　年龄可视化分析

```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
plt.rcParams['font.sans-serif']=['SimHei']
plt.rcParams['axes.unicode_minus'] = False
font = 30
figa, figb = (25,15),(12,12)
dpi_all = 300
data = pd.read_csv('../tmp/数据预处理.csv', encoding='utf-8-sig')
data.set_index(['顾客 ID'],inplace=True)

# 分析年龄分布
age_bins = range(0, max(data['年龄']) + 11, 10)  # 以 10 为间隔创建分组范围
age_groups = ['{}~{}'.format(b, b + 9) for b in age_bins[:-1]]  # 创建分组标签
# 分组并标记分组结果
data['年龄分组'] = pd.cut(data['年龄'], bins=age_bins, labels=age_groups, right=False)
# 计算分组数量并按照分组标签排序
age_distribution = data['年龄分组'].value_counts().sort_index()
age_distribution
# 绘制柱状图
plt.figure(figsize=figa, dpi=dpi_all)
```

```
bars = plt.bar(age_groups, age_distribution)
plt.bar_label(bars, fontsize=font)
plt.xticks(fontsize=font)
plt.yticks(fontsize=font)
plt.xlabel('年龄分布',fontsize=font)
plt.ylabel('人数/人',fontsize=font)
plt.title('年龄分布柱状图',fontsize=font)
plt.tight_layout()  # tight_layout 方法可以保证图像的完整度
plt.show()
# 剔除大于 110 岁的年龄
data = data[data['年龄']<110]
```

结果如图 3-2 所示，本数据集中的顾客年龄主要分布在 20 ~ 69 岁范围内，而且可以观察到以下特征。

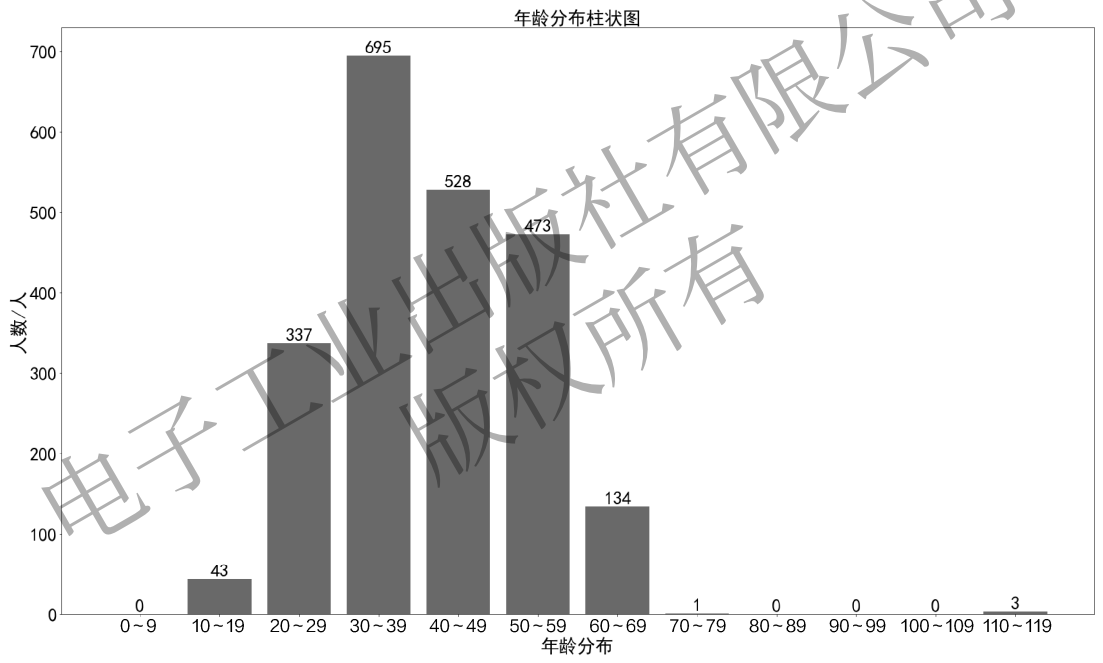

图 3-2　年龄分布柱状图

（1）30 ~ 39 岁的顾客数量最多。这个年龄段的顾客是商超的主要顾客群体。他们可能处于事业发展、家庭成立或养育子女的阶段，在购物方面有一定的消费能力和需求。因此，针对这一年龄段的顾客进行市场推广和产品定位会有更大的潜力。

（2）年龄超过 70 岁的顾客较少。数据显示，只有 4 个顾客年龄超过 70 岁，其中有 3 个顾客的年龄段为 110 ~ 119 岁。但是在 80 ~ 109 岁的年龄段中顾客数量为 0，考虑到这个年龄段的老人不太可能成为新注册顾客，因此认为年龄超过 110 岁的 3 个顾客为异常值，需要对其进行清理。

2）年收入

绘制条形图分析商超顾客的年收入情况，如代码 3-4 所示。

代码 3-4　年收入可视化分析

```
# 年收入分布
data['年收入/元'] = data['年收入/元'].astype('int')
n = 30000
income_bins = range(0, max(data['年收入/元']) + n + 1, n)  # 以 30000 为间隔创建分组范围
income_groups = ['{}~{}'.format(b, b + n) for b in income_bins[:-1]]  # 创建分组标签
# 分组并标记分组结果
data['年收入分组'] = pd.cut(data['年收入/元'], bins=income_bins, labels=income_groups,
right=False)
# 计算分组数量并按照分组标签排序
income_distribution = data['年收入分组'].value_counts().sort_index()
# 绘制条形图
plt.figure(figsize=figa, dpi=dpi_all)
bars = plt.barh(income_groups, income_distribution)
plt.bar_label(bars, fontsize=font)
plt.xticks(fontsize=font)
plt.yticks(fontsize=font)
plt.xlabel('人数/人',fontsize=font)
plt.ylabel('年收入/元',fontsize=font)
plt.title('年收入分布条形图',fontsize=font)
plt.tight_layout()
plt.show()
```

结果如图 3-3 所示，可以观察到以下特征。

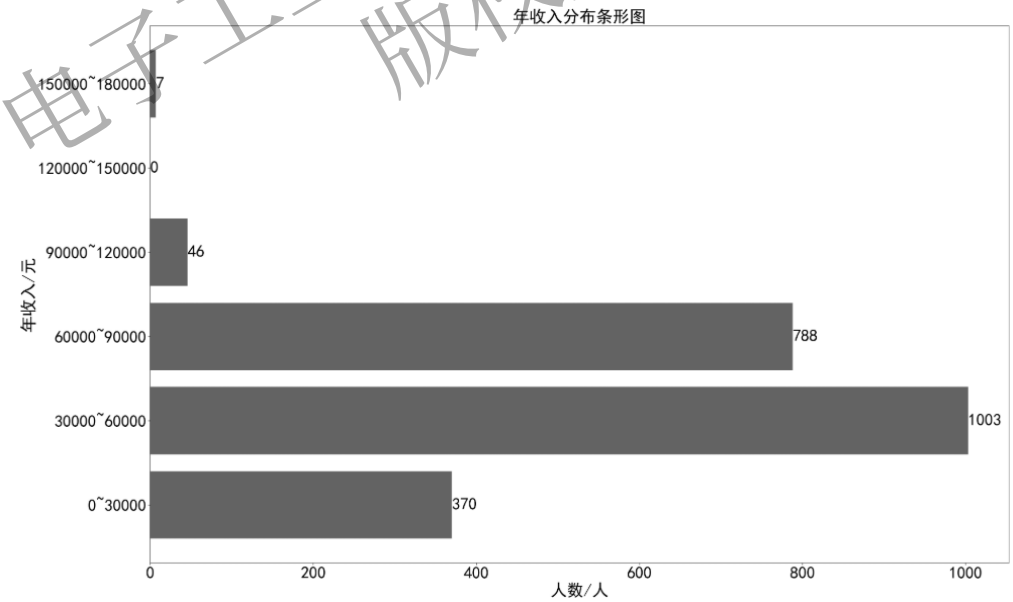

图 3-3　年收入分布条形图

（1）年收入主要集中在 3 万元 ~ 9 万元之间。这个收入范围的顾客数量最多，即该商超主要服务的顾客群体的收入水平属于中等偏上。这可能意味着商超提供的产品或服务定位较高

端，价格相对亲民，符合目标顾客的消费能力和需求。

（2）年收入低于 3 万元或高于 9 万元的顾客人数较少，分别为 370 人和 53 人。这些顾客可能不是商超的主要目标顾客群体，但是也需要考虑他们的消费需求和购物习惯。

3）受教育程度

绘制圆环图分析商超顾客的受教育程度情况，如代码 3-5 所示。

代码 3-5　受教育程度可视化分析

```
# 顾客受教育程度分布
education = data['受教育程度'].value_counts()
plt.figure(figsize=figa, dpi=dpi_all)
colors = plt.get_cmap('BuPu')(np.linspace(0.2, 0.7, len(education)))
patches, texts, autotexts = plt.pie(education.values, labels=education.index,
                                    colors=colors, autopct='%1.1f%%',
                                    startangle=60, wedgeprops=dict(width=0.3,
edgecolor='white'))
plt.setp(autotexts, fontsize=font)
plt.setp(texts, fontsize=font)
plt.axis('equal')
plt.title('受教育程度分布圆环图',fontsize=font)
plt.tight_layout()
plt.show()
```

结果如图 3-4 所示，可以观察到以下特征。

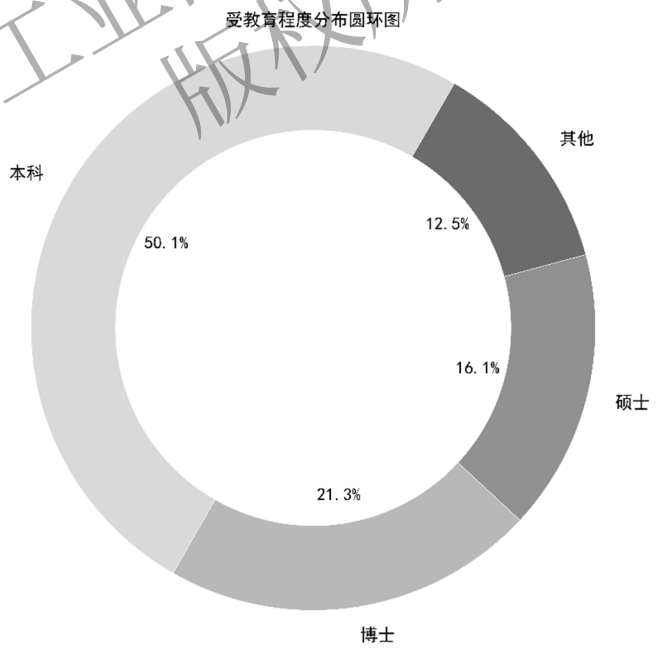

图 3-4　受教育程度分布圆环图

（1）受教育程度为本科的顾客占总顾客数的 50.1%，占比最高。半数顾客的受教育程度为

本科，这意味着这些顾客具有较高的消费能力和购买决策能力，对于商超提供的产品或服务有一定的要求。

（2）受教育程度为硕士和博士的顾客分别占总顾客数的 16.1%和 21.3%。这些顾客可能更加注重商品的品质、品牌和服务质量等方面，因此商超需要针对这些顾客实行更加精准的市场策略和产品定位。

（3）受教育程度为其他的顾客占总顾客数的 12.5%，这些顾客可能具有不同的消费需求和购物习惯，商超需要根据不同顾客的特点，开展不同的营销活动和服务策略。

基于这些数据，商超可以考虑针对不同受教育程度的顾客制定不同的市场推广和产品定位策略，以满足不同顾客的消费需求。例如，可以针对本科及以上的顾客推出高品质、高价值的商品和服务，而对于其他顾客则可以推出更加实惠、适用的商品和服务。

4）婚姻状况

绘制柱状图分析商超顾客的婚姻状况，如代码 3-6 所示。

代码 3-6　婚姻状况可视化分析

```
# 婚姻状况
martial_status = data['婚姻状况'].value_counts()
# 绘制柱状图
categories = martial_status.index
values = martial_status.values
# 计算总和
total = sum(values)
# 计算每个类别的百分比
percentages = [(value/total)*100 for value in values]
# 绘制柱状图
plt.figure(figsize=figa, dpi=dpi_all)
plt.bar(categories, percentages, color='skyblue')
# 显示百分比标签
for i, percentage in enumerate(percentages):
    plt.text(i, percentage + 1, f'{percentage:.1f}%', ha='center',
fontsize=font)
# 添加标题和标签
plt.xticks(fontsize=font)
plt.yticks(fontsize=font)
plt.ylabel('人数占比', fontsize=font)
plt.title('婚姻状况分布柱状图',fontsize=font)
plt.tight_layout()
plt.show()
```

结果如图 3-5 所示，可以观察到以下特征。

（1）已婚顾客占总顾客数的 61.8%，占比最高。在商超主要服务的顾客群体中，大多数顾客是已婚人士。已婚顾客可能更注重日常生活用品、家庭用品及孩子的教育用品等产品的购买。

（2）未婚顾客占总顾客数的 24.5%，占比适中。这部分顾客可能是年轻人或单身人士，他

们可能更加关注时尚、个性化的产品和服务，而且在消费习惯上可能更加自由和多样化。

（3）离异顾客占总顾客数的 10.3%，丧偶顾客占总顾客数的 3.4%，占比较低。这部分顾客可能在日常用品和心理健康产品上有更多的消费需求。

基于这些数据，商超可以针对不同的顾客群体制定相应的市场推广和产品定位策略。例如，可以为已婚顾客提供家庭生活用品、亲子教育用品等；为未婚顾客提供时尚、个性化的商品和服务；为离异和丧偶顾客提供特殊的支持和关怀。

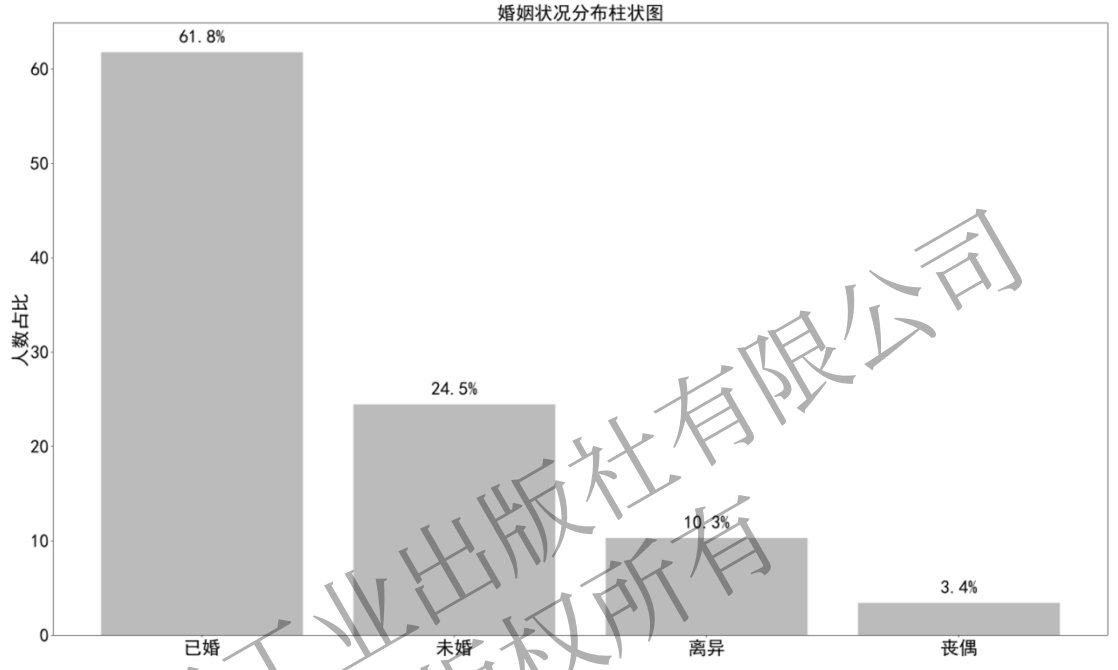

图 3-5　婚姻状况分布柱状图

2. 消费行为

分析顾客的消费行为包括消费频次、消费金额、消费产品类别偏好等，以了解顾客的消费习惯和消费需求，为精准营销和个性化推荐提供依据。

1）品类消费情况

绘制柱状图分析商超顾客的品类消费情况，查看各品类的消费金额，如代码 3-7 所示。

代码 3-7　品类消费情况可视化分析

```
# 品类消费金额
category = data.iloc[:,8:14].sum(axis=0)
# 绘制柱状图
plt.figure(figsize=figa, dpi=dpi_all)
bars = plt.bar(category.index, category.values)
plt.bar_label(bars, fontsize=font)
plt.xticks(fontsize=font)
plt.yticks(fontsize=font)
plt.xlabel('消费品类',fontsize=font)
plt.ylabel('消费金额/元',fontsize=font)
```

```
plt.title('品类消费柱状图',fontsize=font)
plt.tight_layout()
plt.show()
```

结果如图 3-6 所示，可以观察到以下特征。

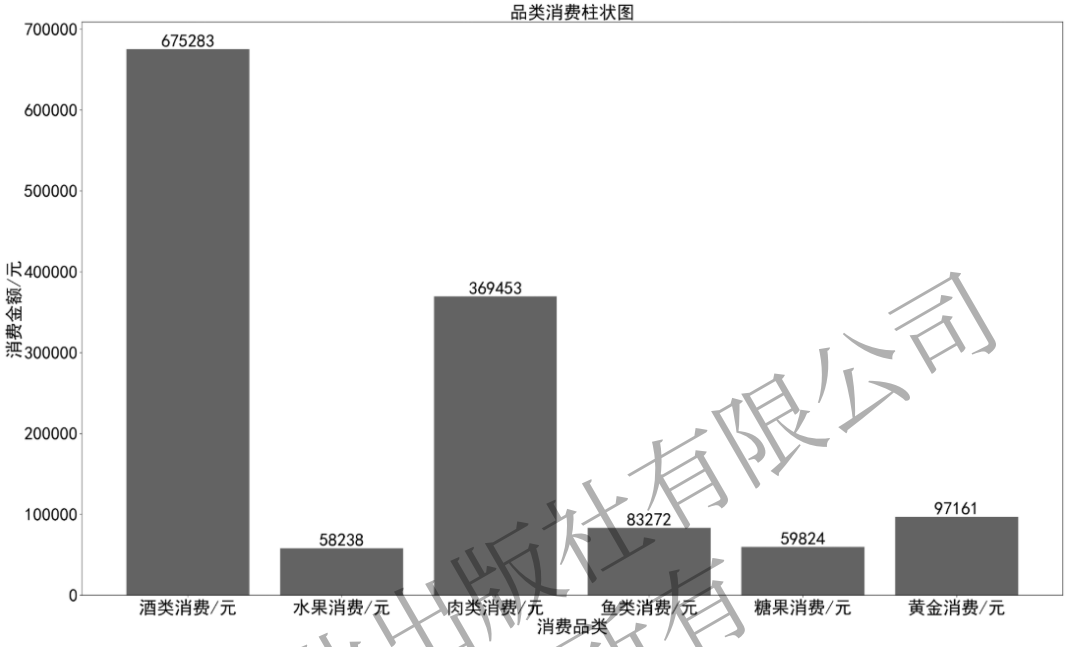

图 3-6　品类消费柱状图

（1）酒类总消费金额最高，达到了 675283 元，这表明商超顾客在购买酒类产品上投入了相当大的金额。有多种原因可能导致这一结果，如酒类产品单价通常较高，且在社交场合、节假日和庆祝活动中往往扮演着重要角色，因此顾客在商超中选择购买酒类产品来满足他们的需求。

（2）肉类总消费金额次之，为 369453 元。这表明商超顾客在购买各种肉类产品时也有相当大的支出。肉类在人们的膳食中通常占据重要地位，因此这一结果并不令人意外。商超可能提供了丰富的肉类选择，满足了顾客对于肉类品质和种类的需求。

（3）水果总消费金额为 58238 元，尽管相对较低，但仍然表明商超顾客在购买水果方面也有一定的花费。水果在健康饮食中起着重要作用，并且受到许多人青睐，因此商超提供各种新鲜和优质的水果可能是吸引顾客购买的原因之一。

（4）鱼类总消费金额为 83272 元，与水果一样，尽管相对较低，但仍然表明商超顾客在购买鱼类方面有一定的花费。鱼类是许多人膳食中重要的优质蛋白来源之一，而且具有丰富的营养价值，因此顾客会选择在商超购买鱼类产品来满足他们的需求。

（5）糖果总消费金额为 59824 元，与水果、鱼类一样，尽管相对较低，但仍然表明商超顾客也会花费一定的金额购买糖果产品。糖果往往是人们日常生活中的休闲零食，也是礼物或节庆活动的一部分，因此商超提供了各种糖果选择，满足了顾客对于糖果的需求。

（6）黄金总消费金额为 97161 元，虽然较酒类、肉类相差较多，但比水果、鱼类和糖果消费相对较高。表明商超顾客在购买黄金上会花费一定的金额。黄金被视为一种重要的资产和投

资，具有保值和增值的特性，因此一些顾客选择在商超购买黄金来满足他们的需求。

2）参与促销情况

分析商超顾客参与促销的情况，如代码 3-8 所示。

代码 3-8　参与促销情况可视化分析

```
# 参与促销人数
category = data.iloc[:,19:25].sum(axis=0)
# 绘制柱状图
plt.figure(figsize=figa, dpi=dpi_all)
bars = plt.bar(category.index, category.values)
plt.bar_label(bars, fontsize=font)
plt.xticks(fontsize=font)
plt.yticks(fontsize=font)
plt.ylabel('参与人数/人',fontsize=font)
plt.title('参与促销人数柱状图',fontsize=font)
plt.tight_layout()
plt.show()
```

结果如图 3-7 所示，可以观察到以下特征。

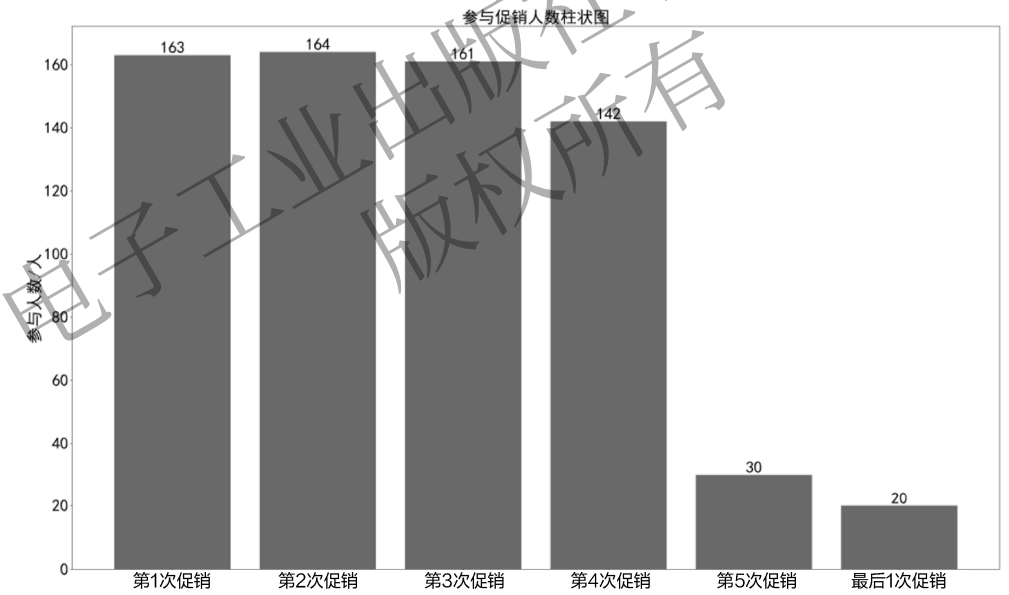

图 3-7　参与促销人数柱状图

（1）从第一次到最近一次的促销活动，参与人数总体上呈现下降趋势。具体来看，参与人数分别为 163 人、164 人、161 人、142 人、30 人和 20 人，其中第 5 次促销活动的参与人数急剧下降。这种情况可能表明商超在前 4 次促销中没有达到顾客期望的优惠幅度或产品选择不够吸引人，导致顾客逐渐失去参与活动的兴趣。

（2）最后一次促销的参与人数明显较少，为 20 人。可能有多种原因导致参与人数如此之少，如缺乏有效的宣传、促销活动吸引力不足、顾客疲劳感等。

商超需要进一步研究和分析促销活动的具体细节和效果，以确定如何在未来的促销活动中吸引更多的顾客参与，提高销售量，这可能包括改进促销策略、增加产品选择、提高促销幅度等因素。

3. 消费渠道

绘制饼状图，对顾客渠道来源进行解析，了解消费者是如何定位至商超并实施购物行为的，为渠道资源配置及推广策略的制定提供决策依据，如代码 3-9 所示。

代码 3-9　消费渠道可视化分析

```
# 消费渠道
category = data.iloc[:,15:18].sum(axis=0)
plt.figure(figsize=figa, dpi=dpi_all)
patches, texts, autotexts = plt.pie(category.values, labels=category.index,
autopct='%1.1f%%',
                                    startangle=90, wedgeprops={'edgecolor': 'white'})
plt.setp(autotexts, fontsize=font)
plt.setp(texts, fontsize=font)
plt.axis('equal')
plt.title('消费渠道人次占比饼图',fontsize=font)
plt.tight_layout()
plt.show()
```

结果如图 3-8 所示，可以观察到以下特征。

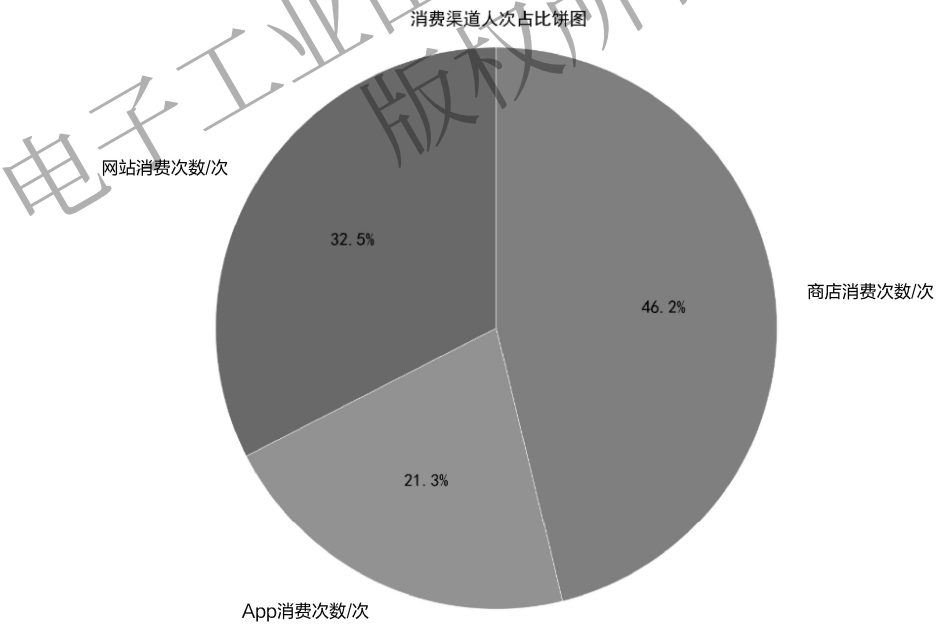

图 3-8　消费渠道人次占比饼图

（1）商店消费的人次占比为 46.2%，显示了线下实体店在顾客消费渠道中的重要性。商超可以进一步加强线下实体店的宣传和推广，提供舒适的购物环境和个性化的服务，以吸引更多顾客到店消费。

（2）网站是顾客找到商超并进行购物的主要渠道之一，占比为 32.5%。这意味着商超的网站对于吸引顾客和促进购物起着重要作用。商超可以进一步优化网站的顾客体验，提高网站的可用性和吸引力，以吸引更多顾客访问和购物。

（3）App 是另一个重要渠道，占比为 21.3%。这表明商超的移动应用程序在顾客购物过程中起到了一定的作用。商超可以投入更多资源来开发和推广移动应用程序，提供更好的顾客体验和功能，以吸引更多顾客使用 App 进行购物。

（4）此外，线上消费渠道（网站和 App）的总占比为 53.8%，明显高于线下消费渠道的 46.2%。这表明线上消费渠道在顾客购物行为中扮演着重要角色，并且具有较大的发展潜力。基于这一情况，商店可以采取一些策略来进一步发展线上消费渠道，如提升线上消费渠道的顾客体验、开展线上独家促销活动、加强线上营销推广、强化线上顾客服务等措施，提升线上消费渠道在整体销售中的占比，同时满足消费者对线上购物的需求，实现线上线下消费渠道的良性互动与发展。

4. 顾客满意度

基于顾客反馈、评价和投诉等数据，绘制饼图分析顾客的满意度和体验感受，了解顾客对商超的整体满意度和改进需求，如代码 3-10 所示。

代码 3-10　顾客满意度可视化分析

```
# 顾客满意度
complain = data['是否投诉'].value_counts()
plt.figure(figsize=figa, dpi=dpi_all)
patches, texts, autotexts = plt.pie(complain.values, labels=['无投诉','有投诉'],
autopct='%1.1f%%',
                                    startangle=90, wedgeprops={'edgecolor': 'white'})
plt.setp(autotexts, fontsize=font)
plt.setp(texts, fontsize=font)
plt.axis('equal')
plt.title('投诉情况占比饼图',fontsize=font)
plt.tight_layout()
plt.show()
```

结果如图 3-9 所示，可以观察到以下特征。

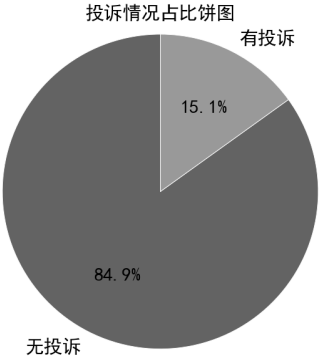

图 3-9　投诉情况占比饼图

（1）无投诉占比为 84.9%，说明近两年大部分顾客对商超的整体满意度较高。这表明商超在服务质量、产品质量、售后支持等方面取得了一定的成就，能够满足大部分顾客的需求。

（2）有投诉占比为 15.1%，虽然相对较低，但表示仍然存在一部分顾客对商超的服务不满意或遇到了问题。商超需要重视这部分顾客的反馈和投诉，并积极采取措施进行改进，提高顾客的满意度，并提升顾客的忠诚度和口碑效应。

（3）此外，有一部分顾客可能在遇到问题或不满意的情况下选择默默离开，而并未进行投诉或表达意见。这类顾客虽无法通过数据直接反映出来，但他们的离开对商超的业务发展同

样具有一定的影响。商超可以通过调查问卷、顾客调研等方式获取更多的顾客反馈，同时可以通过社交媒体、在线评论等途径收集顾客的意见和建议，积极改善相应问题，提升顾客满意度和忠诚度，减少顾客无声离开的情况发生。

3.1.3 相关性分析

顾客的不同信息之间存在一定的相关性，针对数值型属性计算属性之间的相关性并绘制热力图，以发现不同属性之间的关联程度，揭示潜在的影响因素和规律，为后续的建模指标选取提供依据。

1. 计算相关性

选择数值型属性作为相关性分析的对象，使用皮尔逊相关系数计算每对属性之间的相关性。相关系数的取值范围为−1 ~ 1，其中 1 表示完全正相关，−1 表示完全负相关，0 表示无相关性。

绘制热力图并使用不同的颜色来代表相关系数，展示属性之间的相关性。热力图中的方格代表两个属性之间的相关性，颜色越深表示相关性越强。通过热力图可以直观地看到属性之间的相关性程度，如代码 3-11 所示。

代码 3-11 相关性分析

```
# 相关性分析
import seaborn as sns
import matplotlib.pyplot as plt
from matplotlib.colors import LinearSegmentedColormap
import pandas as pd

data = pd.read_csv('../tmp/数据预处理.csv', encoding='utf-8-sig')
data.set_index(['顾客 ID'],inplace=True)
# 计算相关系数矩阵
data.columns
corcols = ['年龄','年收入/元','注册天数/天','距上次消费天数/天',
       '酒类消费/元','水果消费/元','肉类消费/元','鱼类消费/元','糖果消费/元','黄金消费/元',
       '折扣消费次数/次','网站消费次数/次','App 消费次数/次','商店消费次数/次',
       '上月网站访问次数/次']
corr_matrix = data[corcols].corr()
# 绘制热力图矩阵
# 定义 colormap 的颜色列表
colors = [(1, 1, 1), (0, 0, 0.5)]
# 创建 colormap 对象
cmap = LinearSegmentedColormap.from_list('my_cmap', colors)
plt.figure(figsize=(15,10), dpi=300) # 设置绘图尺寸和每英寸点数
plt.rcParams['font.sans-serif']=['SimHei']
plt.rcParams['axes.unicode_minus'] = False
sns.heatmap(corr_matrix, cmap=cmap, annot=False, fmt=".2f", annot_kws={"size": 20})
plt.tick_params(axis='both', which='major', labelsize=15)
```

```
# 显示图形
plt.tight_layout()
plt.show()
```

结果如图 3-10 所示，可以观察到以下特征。

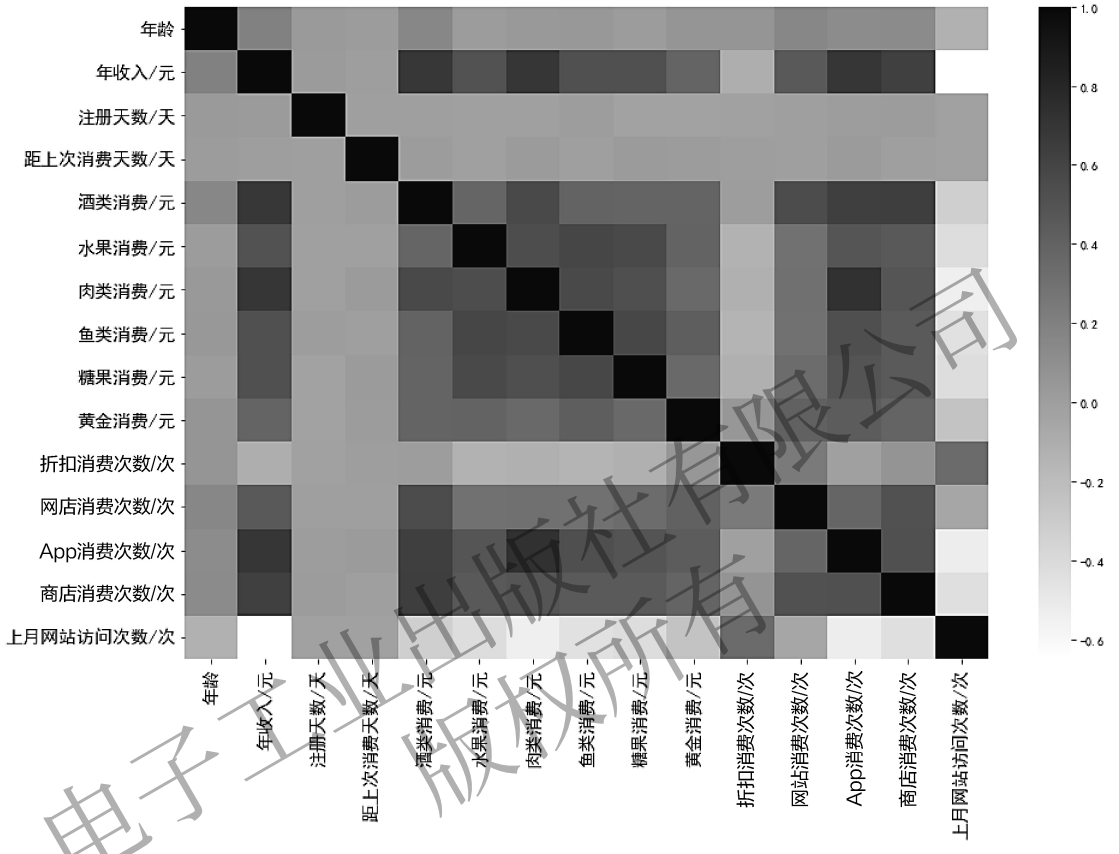

图 3-10　相关性热力图

（1）年龄和年收入之间的相关性热力图显示颜色较浅，相关系数为 0.1997，表明两者存在一定程度的正相关性，但相关性并不十分显著。

（2）注册天数与其他特征之间的相关性热力图显示颜色均较浅，绝对值均不超过 0.1，因此可以认为注册天数与其他特征之间的线性相关性并不明显。

（3）针对消费类别观察，酒类消费与肉类消费、鱼类消费、糖果消费之间的相关性热力图呈现较深的颜色，表明存在一定程度的正相关性，这可能意味着这些消费项目之间存在某种购买偏好或消费行为之间的关联。另外，水果消费也与其他消费类别呈现一定的正相关性，不过相关性程度相对较低。

（4）折扣消费次数与各项消费之间的相关性热力图显示颜色较浅，这说明折扣消费次数与消费项目之间的线性相关性不显著。

（5）网站消费次数、App 消费次数、商店消费次数与各项消费之间的相关性热力图显示颜色较深，表明相关性较为显著且呈现出较强的正相关性，这说明购买方式与消费项目之间存在一定程度的关联。

2. 属性构造

选取热力图中颜色较深、相关性较强的属性进行综合，构造出新的属性。选取酒类消费、水果消费、肉类消费、鱼类消费、糖果消费、黄金消费这 6 个属性进行综合，用于构造消费金额属性；选取网站消费次数、App 消费次数、商店消费次数这 3 个属性进行综合，用于构造折扣消费比例，具体的属性构造过程将在下一任务中介绍。

任务 3.2　构建聚类模型并分析结果


构建聚类模型

商超顾客价值分析模型的构建与结果的分析主要包括以下 3 部分。

（1）基于 RFM 模型，筛选和构造合适的指标，建立用于聚类的 RFMPI 模型。

（2）根据商超顾客的 5 个指标，对顾客进行聚类分组。通过聚类分析将顾客划分为不同的群体，从而识别顾客群体之间的相似性和差异性，为后续的顾客价值分析提供基础。

（3）结合业务需求对每个顾客群体进行特征分析，分析不同顾客群体的顾客价值，并对顾客群体进行排名。深入分析每个顾客群体的购买行为、消费偏好、忠诚度等数据，评估各顾客群体的潜在商业价值，为商超制定个性化营销策略和服务方案提供依据。

3.2.1　筛选与构造建模指标

RFM 模型通过最近一次消费时间、消费频率和消费金额来划分顾客群体，但存在一定的局限性。引入更多指标可以改进 RFM 模型，如顾客基本情况、顾客消费偏好等，从而更为全面地实现顾客价值分析。合理选择指标并结合改进 RFM 模型，可以帮助商超更好地了解顾客、优化营销策略，实现顾客价值最大化。

1. RFM 模型

RFM 模型是一种用于顾客分群分析的常用方法，RFM 模型基于 3 个关键指标，即 Recency（最近一次消费）、Frequency（消费频率）和 Monetary（消费金额）。将顾客按照这 3 个指标进行分组，可以更好地了解不同类型顾客的行为和价值。

Recency（最近一次消费）：该指标用于衡量顾客最近一次消费产品或服务的时间，通常以天或月为单位。较短的 Recency 值表示顾客近期有消费行为，可能对产品或服务更感兴趣。

Frequency（消费频率）：该指标用于衡量顾客在一定时间范围内消费产品或服务的次数。频繁消费的顾客往往具有更高的忠诚度，并且可能对交叉销售和升级产品感兴趣。

Monetary（消费金额）：该指标用于衡量顾客在一定时间范围内消费产品或服务的总金额。较高的 Monetary 值表示顾客在消费时倾向于花费更多的资金，并且可能是高价值顾客。

在 RFM 模型中，通常将每个指标分成几个等级或分组，如高、中、低，根据这些等级或分组可以形成一个 3D 的顾客分析空间。通过在这个空间中对顾客进行定位，可以得到不同群体的顾客特征和行为模式。

RFM 模型的主要应用包括以下 3 点。

（1）顾客细分：将顾客分成不同的群体，如重要价值顾客、沉睡顾客、新顾客等，以便更好地针对不同群体制定营销策略。

（2）交叉销售和升级：根据顾客的 RFM 指标，推荐相关产品或服务，提高顾客的消费频率和消费金额。

（3）顾客保持和回流：通过了解顾客的 Recency 值，判断哪些顾客处于流失风险，并采取相应措施（如个性化促销、顾客关怀等）来保持和回流顾客。

2. RFMPI 模型

RFM 模型是一个简单而有效的顾客分析方法，可以帮助商超更好地理解顾客的行为和价值，但是 RFM 模型也有一些缺点，它可能忽略一些重要因素，如顾客消费行为、基本情况等。因此，需要改进 RFM 模型，使其更加精细化、个性化，从而更好地满足实际需求。将顾客的年收入和折扣消费比例作为指标加入 RFM 模型可以进一步丰富模型，从而更为全面地分析顾客的行为和价值。

年收入：将顾客的年收入作为一个附加维度，可以帮助商超了解顾客的经济实力和消费能力。

折扣消费比例：折扣消费比例表示顾客使用折扣消费的次数占总消费次数的比例，该指标揭示了顾客对价格的敏感程度和对优惠活动的响应情况。折扣消费比例的计算方法如下。

$$\text{折扣消费比例}=\frac{\text{折扣消费次数}}{\text{网站消费次数}+\text{App消费次数}+\text{商店消费次数}}$$

将顾客消费时间 R、消费频率 F、消费金额 M、折扣消费比例 P、年收入 I 五个特征作为商超顾客价值分析的主要指标，记为 RFMPI 模型，如代码 3-12 所示。

代码 3-12　筛选与构建指标

```
# 筛选指标，构建 RFMPI 模型
import pandas as pd
data = pd.read_csv('../tmp/数据预处理.csv', encoding='utf-8-sig')
data.set_index(['顾客 ID'],inplace=True)
# 构建指标
data['消费频率'] = data['网站消费次数/次'] + data['App 消费次数/次'] + data['商店消费
次数/次']
data['消费金额'] = data['酒类消费/元'] + data['水果消费/元'] + data['肉类消费/元'] +
data['鱼类消费/元'] + data['糖果消费/元'] + data['黄金消费/元']
data['儿童数量'] = data['儿童数量/人'] + data['青少年数量/人']
data['促销次数'] = data['第 1 次促销'] + data['第 2 次促销'] + data['第 3 次促销'] +
data['第 4 次促销'] + data['第 5 次促销'] + data['最近一次促销']

def calculate_discount_ratio(row):
    if row['消费频率'] != 0:
        return row['折扣购买次数/次'] / row['消费频率']
    else:
        return 0

data['折扣比例'] = data.apply(calculate_discount_ratio, axis=1)
data = data[data['折扣比例']<=1]
RFMPI = data[['距上次消费天数/天','消费频率','消费金额','折扣比例','年收入/元']]
RFMPI.columns = ['消费时间', '消费频率', '消费金额','折扣比例','年收入']
# 指标排名
```

```
rp_labels = range(4, 0, -1)
fm_labels = range(1,5)
r_quartiles = pd.qcut(RFMPI['消费时间'], 4, labels = rp_labels)
RFMPI = RFMPI.assign(R = r_quartiles.values)
f_quartiles = pd.qcut(RFMPI['消费频率'], 4, labels = fm_labels)
RFMPI = RFMPI.assign(F = f_quartiles.values)
m_quartiles = pd.qcut(RFMPI['消费金额'], 4, labels = fm_labels)
RFMPI = RFMPI.assign(M = m_quartiles.values)
p_quartiles = pd.qcut(RFMPI['折扣比例'], 4, labels = rp_labels)
RFMPI = RFMPI.assign(D = p_quartiles.values)
def join_RFMPI(x):
    return str(int(x['R'])) + str(int(x['F'])) + str(int(x['M'])) + str(int(x['D']))

RFMPI['RFMPI_Segment'] = RFMPI.apply(join_RFMPI, axis=1)
RFMPI['RFMPI_Score'] = RFMPI[['R','F','M','D']].sum(axis=1)
RFMPI.to_csv('../tmp/RFMPI.csv', encoding='utf-8-sig')
```

3.2.2 构建与训练聚类模型

对商超顾客进行聚类的过程包括以下 4 个步骤。

（1）数据标准化：对选定的特征进行标准化处理，确保各特征在相似的数值范围内，避免因为量纲不一致导致的聚类结果不准确问题。

（2）确定最佳聚类数 K：根据业务需求和数据特点，确定要将顾客分成的簇的数量 K 值。

（3）模型训练：利用 K 均值算法对标准化后的顾客数据进行聚类，根据 K 值初始化质心，并迭代更新质心直至达到停止条件。

（4）聚类结果分析：根据聚类结果将顾客分成不同的簇，每个簇代表一组具有相似消费特征的顾客群体。

1. 确定最佳聚类数

要确定最佳聚类数可以使用肘方法，肘方法是一种常用的、在聚类分析中确定最佳聚类数的方法。它基于观察不同聚类数下的聚类误差平方和（SSE）的变化情况，通过绘制聚类数与 SSE 的曲线，找到一个“肘点”，即曲线开始呈现拐点的位置，确定最佳的聚类数。肘方法包括以下 5 个步骤，具体实现如代码 3-13 所示。

（1）运行聚类分析：使用所选的聚类算法（如 K 均值聚类），将数据集分成不同的聚类数，并计算每个聚类数下的 SSE。

（2）计算 SSE：对于每个聚类数，计算聚类结果中各点到其所属聚类中心的距离平方和，即 SSE。

（3）绘制肘曲线：将聚类数与对应的 SSE 绘制成图表（通常是折线图）。横轴表示聚类数，纵轴表示 SSE。

（4）分析肘点：观察肘曲线，找到一个拐点或肘点。肘点是指曲线开始出现明显减缓的位置，形状类似手臂的肘部。该点表示增加更多的聚类数对 SSE 的改善效果递减，因此可以被认为是最佳聚类数。

（5）选择最佳聚类数：根据观察到的肘点选择相应的聚类数作为最佳聚类数，并根据该聚

类数重新运行聚类算法，得到最终的聚类结果。

代码 3-13　确定最佳聚类数

```
# Kmeans
from sklearn.cluster import KMeans
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif']=['SimHei']
plt.rcParams['axes.unicode_minus'] = False
font = 30
figa, figb = (25,15),(10,10)
RFMPI = pd.read_csv('../tmp/RFMPI.csv', encoding='utf-8-sig')

# 数据标准化
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
data_k0 = RFMPI.iloc[:,:6]
data_k = pd.DataFrame(scaler.fit_transform(data_k0[['消费时间', '消费频率', '消费
金额','折扣比例','年收入']]))
data_k = pd.concat([data_k0['顾客 ID'],data_k],axis=1)
data_k.set_index(['顾客 ID'],inplace=True)

# 使用肘方法确定聚类数
from scipy.spatial.distance import cdist
# 定义聚类数的范围
k_values = range(1, 10)
distortions = []
# 计算每个聚类数对应的聚类误差平方和
for k in k_values:
    kmeans = KMeans(n_clusters=k, n_init=10)
    kmeans.fit(data_k)
    distortions.append(sum(np.min(cdist(data_k, kmeans.cluster_centers_, 'euclidean'),
axis=1)) / data_k.shape[0])
# 绘制肘部曲线
plt.figure(figsize=(15,10), dpi=300)
plt.plot(k_values, distortions, 'bx-')
plt.xlabel('聚类数/个',fontsize=font)
plt.ylabel('聚类误差平方和',fontsize=font)
plt.xticks(fontsize=font)
plt.yticks(fontsize=font)
plt.title('肘方法',fontsize=font)
plt.tight_layout()
plt.show()
```

运行代码 3-13，得到 SSE 随聚类数的变化曲线，如图 3-11 所示。

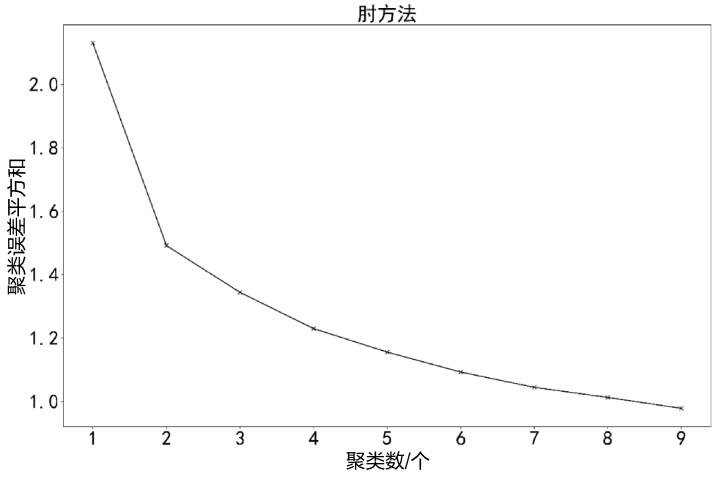

图 3-11　SSE 随聚类数的变化曲线

由图 3-11 可以看出，当聚类数 K 取 4 时，曲线开始出现明显减缓，即增加更多聚类数对 SSE 的改善效果已经递减。因此，取 4 为最佳聚类数。

2. 创建与训练 K-Means 聚类模型

根据选取的 RFMPI 模型，利用 K-Means 聚类模型将数据聚合为 4 个不同的类别，如代码 3-14 所示。

代码 3-14　创建与训练 K-Means 聚类模型

```
# 使用 K-Means 聚类模型
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
# 设置聚类的簇数
k = 4
# 创建 K-Means 聚类模型并进行训练
kmeans = KMeans(n_clusters=k, random_state=1234, n_init=10).fit(data_k)
# 获取每个样本所属的簇标签
labels = kmeans.labels_
# 将簇标签添加到数据集中
data_k['类别'] = labels
data_k.columns = ['消费时间', '消费频率', '消费金额','折扣比例','年收入','类别']
```

3.2.3　聚类结果分析

通过可视化分析聚类结果中各顾客群体的顾客特征、人数占比、年龄分布和年收入分布。

1. 分析各顾客群体的顾客特征

绘制雷达图对分群结果进行分析，分析各顾客群体的顾客特征，如代码 3-15 所示。

代码 3-15　分析各顾客群体的顾客特征

```
# 结果分析
```

```
# 绘制雷达图
cluster_labels = [0, 1, 2, 3]  # 聚类标签
cluster_colors = ['red', 'green', 'blue','yellow']  # 聚类颜色
cluster_means = {}
for label in cluster_labels:
    cluster_means[label] = data_k[data_k['类别'] == label].iloc[:,:4].mean(axis=0)

cluster_means = data_k.groupby(['类别']).mean().transpose()
fig, ax = plt.subplots(figsize=(6,6), subplot_kw=dict(polar=True), dpi=300)
angles = np.linspace(0, 2 * np.pi, 5, endpoint=False).tolist()
angles += angles[:1]
cluster_colors = ['red', 'green', 'blue','orange']
line_styles = ['-', '--', ':','-.']
for label in cluster_labels:
    values = cluster_means.iloc[:,label].tolist()
    values += values[:1]
    line, = ax.plot(angles, values, linewidth=2, label=f'Cluster {label}',
           color=cluster_colors[label], linestyle=line_styles[label])
    # ax.fill(angles, values, alpha=0.3, color=line.get_color())

ax.set_thetagrids(np.degrees(angles[:-1]), labels=['消费时间', '消费频率', '消费
金额','折扣消费比例','年收入'])
ax.legend(labels = ['类别 0', '类别 1', '类别 2', '类别 3'])
plt.tight_layout()
plt.show()
```

运行代码 3-15，得到顾客特征雷达图，如图 3-12 所示，需要注意每次聚类后的类别结果都有可能发生变动。

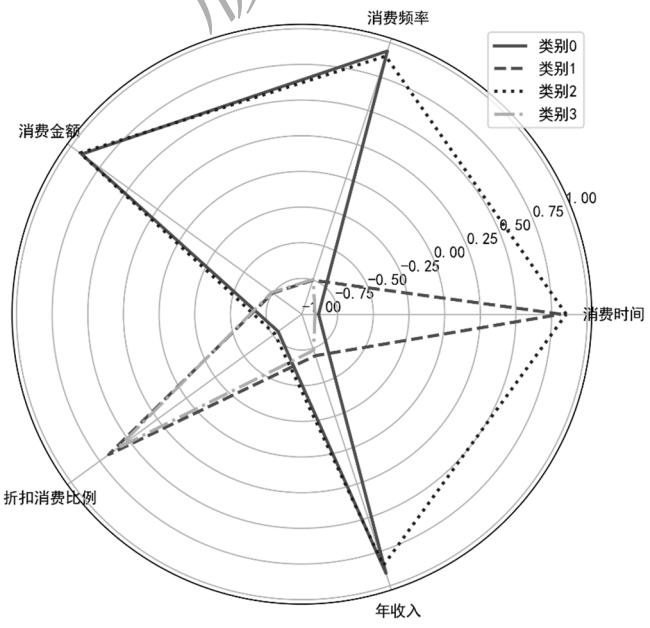

图 3-12　顾客特征雷达图

由图 3-12 可以看出，不同顾客群体的顾客特征如表 3-3 所示。

表 3-3　不同顾客群体的顾客特征

顾客群体	顾客特征
类别 0	该类别中的顾客消费时间（最近一次消费距今时间）较短、消费频率较高、消费金额较高、折扣消费比例（使用折扣消费次数占比）较低、年收入较高。这可能代表了一组高价值顾客或忠实顾客，该类顾客的消费频率较高，消费金额较大，对折扣比例不太敏感，且年收入较高
类别 1	该类别中的顾客消费时间较长、消费频率较低、消费金额较低、折扣消费比例较高、年收入较低。这可能代表了一组比较谨慎或预算有限的消费者。该类顾客倾向于使用折扣进行消费，消费金额相对较小，消费频率也不太高。年收入较低可能是该类顾客消费能力的一个限制因素。消费时间较长表明这组顾客的流失风险较大
类别 2	该类别中的顾客消费时间较长、消费频率较高、消费金额较高、折扣消费比例较低、年收入较高。这个类别与类别 0 相似，也可能代表了一组高价值顾客或忠实顾客。该类顾客消费频率相对较高，消费金额也较大，表明该类顾客对高价值产品感兴趣。该类顾客的年收入较高，说明该类顾客有更多的消费能力。但消费时间较长，表明这组顾客可能有流失的风险
类别 3	该类别中的顾客消费时间较短、消费频率较低、消费金额较低、使用折扣消费次数占比较高、年收入较低。这个类别与类别 1 相似，但消费时间较近。这可能代表了一组潜在顾客。该类顾客消费频率较低，消费金额较小，对折扣比例敏感，并且年收入较低

2. 分析各顾客群体的人数占比

绘制饼图分析各顾客群体的人数占比情况，如代码 3-16 所示。

代码 3-16　分析各顾客群体的人数占比情况

```
# 各顾客群体人数
data = pd.read_csv('../tmp/数据预处理.csv', encoding='utf-8-sig')
data.set_index(['顾客 ID'],inplace=True)
data = data[data['年龄']<110]
data = pd.concat([data,data_k['类别']],axis=1)
data = data.dropna()
data['类别'] = data['类别'].astype(int)
category = data['类别'].value_counts()
plt.figure(figsize=figb, dpi=300)
patches, texts, autotexts = plt.pie(category.values, labels=category.index,
autopct='%1.1f%%', startangle=90, wedgeprops={'edgecolor': 'white'})
plt.setp(autotexts, fontsize=font)
plt.setp(texts, fontsize=font)
plt.axis('equal')
plt.title('各顾客群体人数占比饼图',fontsize=font)
plt.tight_layout()
plt.show()
```

运行代码 3-16，得到各顾客群体人数占比饼图，如图 3-13 所示。

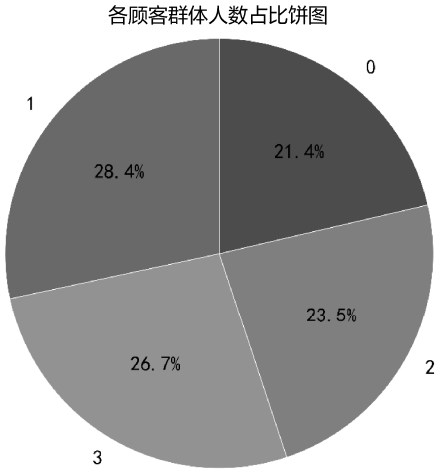

图 3-13　各顾客群体人数占比饼图

由图 3-13 可以看出，各顾客群体的人数相对接近，占比大致相同。这表明在这个样本中，各顾客群体的人数相对均衡，没有明显的数量优势或劣势。

3. 分析各顾客群体的年龄分布

绘制箱线图分析各顾客群体年龄的分布情况，如代码 3-17 所示。

代码 3-17　分析各顾客群体年龄的分布情况

```
# 各顾客群体年龄
# 按照类别进行分组
grouped_data = [data[data['类别'] == category]['年龄'] for category in data['类别'].unique()]
# 创建箱线图
plt.figure(figsize=figa, dpi=300)
plt.boxplot(grouped_data, labels=data['类别'].unique())
# 添加标题和标签
plt.xticks(fontsize=font)
plt.yticks(fontsize=font)
plt.title('各顾客群体年龄箱线图',fontsize=font)
plt.xlabel('类别',fontsize=font)
plt.ylabel('年龄',fontsize=font)
# 显示箱线图
plt.tight_layout()
plt.show()
```

运行代码 3-17，得到各顾客群体年龄箱线图，如图 3-14 所示。

由图 3-14 可以看出，类别 3 和类别 1 的平均年龄相对较低，而类别 2 和类别 0 的平均年龄相对较高，但类别间的平均年龄差距较小，即不同顾客群体的平均年龄和年龄范围相似。这意味着不同类别的顾客在年龄上没有明显的差异。

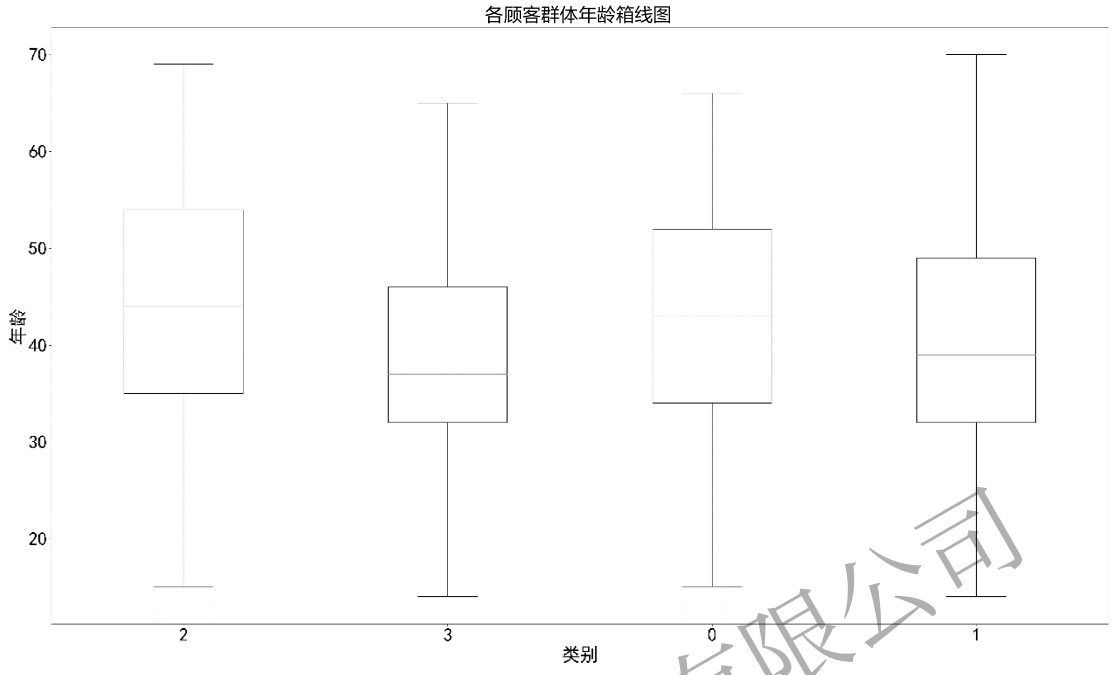

图 3-14　各顾客群体年龄箱线图

4. 分析各顾客群体的年收入分布

绘制箱线图分析各顾客群体的年收入分布情况，如代码 3-18 所示。

代码 3-18　分析各顾客群体的年收入分布情况

```
# 各顾客群体年收入
# 按照类别进行分组
grouped_data = [data[data['类别'] == category]['年收入/元'] for category in
data['类别'].unique()]
# 创建箱线图
plt.figure(figsize=figa, dpi=300)
plt.boxplot(grouped_data, labels=data['类别'].unique())
# 添加标题和标签
plt.xticks(fontsize=font)
plt.yticks(fontsize=font)
plt.title('各顾客群体年收入箱线图',fontsize=font)
plt.xlabel('类别',fontsize=font)
plt.ylabel('年收入/元',fontsize=font)
# 显示箱线图
plt.tight_layout()
plt.show()
```

运行代码 3-18，得到各顾客群体年收入箱线图，如图 3-15 所示。

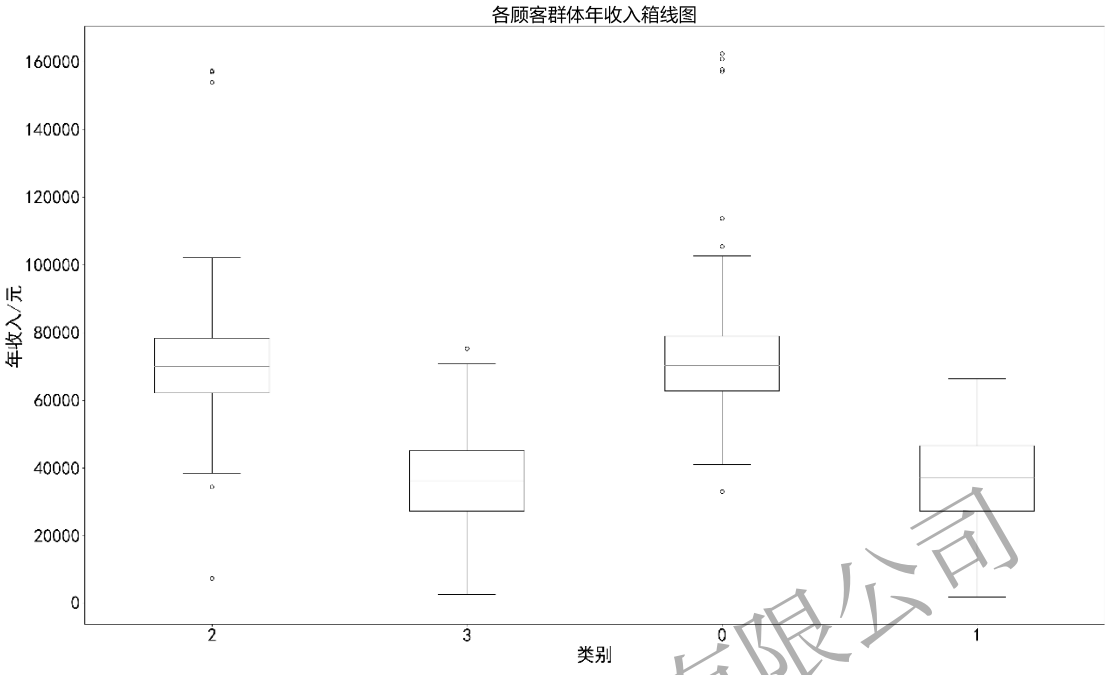

图 3-15　各顾客群体年收入箱线图

由图 3-15 可以看出，不同类别的顾客群体在年收入方面存在一定的差异。类别 2 和类别 0 的顾客群体具有较高的年收入水平，而类别 3 和类别 1 的顾客群体具有较低的年收入水平。

项目总结

本项目的主要目的是通过 RFMPI 模型和 K-Means 聚类算法对顾客进行分群，并通过分析聚类结果分析不同顾客群体的特征。首先通过对选定数据进行探索分析和预处理，利用数据质量评估、可视化分析和相关性分析等方法，确保数据的准确性、完整性。其次，应用 RFMPI 模型和 K-Means 聚类算法进行顾客分群，并利用肘方法确定最佳聚类数，揭示顾客群体中的潜在模式和趋势。最后，针对不同价值的顾客群体展开分析，并将其作为制定个性化营销策略和服务方案的依据，以提升顾客满意度，增强竞争力，实现可持续发展。

实训工单

学　　院		专　　业	
姓　　名		学　　号	
小组成员		组长姓名	

实训目标

1. 对航空客户数据进行探索性分析。
2. 对航空客户进行数据预处理。

3. 使用 K-Means 聚类算法进行客户价值分析。

4. 根据模型聚类结果，分析不同客户群体的特征，撰写工单总结报告。

一、工作任务

随着信息时代的到来，企业营销的焦点已从单纯的以产品为中心逐渐转向以客户为中心，客户关系管理成为企业提升竞争力的核心问题。通过科学的分类方法，企业能够区分高价值客户与低价值客户，针对不同客户群体制定个性化的服务方案，优化资源配置，实现企业利润最大化的目标。然而，面对海量的客户数据和激烈的市场竞争，传统的人工分类方式效率低下，导致部分高价值客户的服务体验未能得到充分优化，影响了企业的市场表现与客户忠诚度。

为此，国内某航空公司决定引入智能化数据分析技术，构建客户价值评估与分类系统。通过整合会员档案信息、航班乘坐记录等多源数据，结合机器学习算法与数据挖掘技术，实现对客户价值的自动评估与精准分类。系统将为不同客户类别提供个性化服务建议，制定差异化的营销策略，帮助航空公司优化资源分配，提高客户满意度与忠诚度，有效提升企业竞争力。航空数据治理团队委派数据分析师小李负责该项目，要求其完成客户分类模型的开发与验证工作，确保技术方案能够精准识别客户价值的核心特征，为航空公司制定营销策略提供可靠的数据支撑，推动客户关系管理能力与服务质量的全面提升。

二、任务准备

1. 数据概况

air_data.csv 文件记录了航空公司系统内的客户基本信息、乘机信息、积分信息等详细数据，属性描述如下表所示。

数据	属性名称	属性说明
客户基本信息	MEMBER_NO	会员卡号
	FFP_DATE	入会时间
	FIRST_FLIGHT_DATE	第一次飞行日期
	GENDER	性别
	FFP_TIER	会员卡级别
	WORK_CITY	工作地所在城市
	WORK_PROVINCE	工作地所在省份
	WORK_COUNTRY	工作地所在国家
	AGE	年龄
乘机信息	FLIGHT_COUNT	观测窗口内的飞行次数
	LOAD_TIME	观测窗口的结束时间
	LAST_TO_END	最后一次乘机时间至观测窗口结束时长
	AVG_DISCOUNT	平均折扣率
	SUM_YR	观测窗口的票价收入
	SEG_KM_SUM	观测窗口的总飞行公里数
	LAST_FLIGHT_DATE	末次飞行日期
	AVG_INTERVAL	平均乘机时间间隔
	MAX_INTERVAL	最大乘机间隔

续表

数据	属性名称	属性说明
积分信息	EXCHANGE_COUNT	积分兑换次数
	EP_SUM	总精英积分
	PROMOPTIVE_SUM	促销积分
	PARTNER_SUM	合作伙伴积分
	POINTS_SUM	总累计积分
	POINT_NOTFLIGHT	非乘机的积分变动次数
	BP_SUM	总基本积分

2. 工具准备

编程语言：Python

工具库：_____

开发 IDE：_____

三、任务实施

1. 对原始数据进行探索性分析。

① 进行描述性统计，查看数据的空值个数、最大值、最小值等信息。

关键代码：_____

② 从客户基本信息、乘机信息、积分信息等方面进行数据探索，寻找客户的分布规律。

关键代码：_____

③ 计算客户信息各属性间的相关性系数矩阵，绘制热力图分析各属性间的相关性。

关键代码：_____

2. 采用数据清洗、属性规约与数据变换等数据预处理方法。

① 进行数据清洗，处理数据中的缺失值、重复值、异常值。

关键代码：_____

② 根据航空公司客户价值 LRFMC 模型，选择与 LRFMC 指标相关的属性。

关键代码：_____

③ 进行数据格式转化、标准化处理。

关键代码：_____

3. 构建客户价值分析模型。

① 使用 K-Means 聚类算法对客户数据进行客户分群。

关键代码：_____

② 根据聚类结果进行特征分析，绘制客户分群雷达图。

关键代码：_____

4. 根据航空公司用户数据分析结果，撰写工作报告。

主要结论：_____

四、评价反馈

根据自己在课堂中的实际表现进行自我反思和自我评价。

自我反思：_____

_____。

自我评价：_____

_____。

评价结论：□优秀　□良好　□一般　□需改进

实训成绩单

评分项目	评分标准	分值	得分
数据探索与分析	能够正确进行数据描述性统计	5	
	能够对用户数据进行数据分布探索分析	10	
	能够计算相关系数矩阵并绘制热力图	10	
数据预处理	能够正确进行数据清洗	5	
	能够根据模型选择相关的属性	10	
	能够正确进行数据转化和标准化	5	
模型建构	能够正确使用 K-Means 聚类算法进行客户分群	10	
	能够根据聚类结果进行特征分析	10	
	能够绘制客户分群雷达图	10	
报告撰写	报告结构完整，逻辑清晰	2	
	分析方法多样，结果准确	2	
	图表清晰，能够有效传达数据信息	2	
	语言简洁，专业术语使用准确	2	
	结论基于数据，建议具体且具有实际应用价值	2	
评价反馈	提交的代码规范	5	
	能对自身进行客观评价	5	
	在实施过程中能发现自身的问题	5	
得分（满分 100 分）			

五、思考与练习

1. 选择题

（1）在数据探索分析中，以下哪种方法用于处理缺失值？（　　）

A. 直接删除所有缺失值的行　　B. 使用均值填充缺失值

C. 使用中位数填充缺失值　　D. 以上都是

（2）以下哪种方法可以用于检测数据中的异常值？（　　）

A. 标准化　　B. 归一化

C. 箱线图　　D. 雷达图

（3）在数据预处理中，以下哪种方法适用于将分类特征转换为数值特征？（　　）

A. 独热编码　　B. 归一化

C. 标准化　　D. 二值化

（4）在 Python 中，以下哪个库可以用于实现 K-Means 聚类算法？（　　）

A. TensorFlow　　B. scikit-learn

C. PyTorch　　D. Pandas

（5）雷达图在聚类结果中的主要作用是什么？（　　）

A. 显示聚类中心的分布　　B. 比较不同聚类的特征分布

C. 展示聚类的层次结构　　D. 可视化聚类的数量

2. 填空题

（1）在 Python 数据探索分析中，查看数据集的基本统计信息通常使用_____函数。

（2）在数据预处理时，处理缺失值的常用方法包括删除、填充或使用_____。

（3）在 Python 数据分析与处理时，数据标准化使用 Scikit-learn 库中的_____类。

（4）在 Python 中，实现 K-Means 聚类算法的库是_____。

（5）使用雷达图可以直观展示不同聚类中心的特征分布，绘制雷达图通常使用_____库。

3. 判断题

（1）数据探索分析的第一步通常是查看数据的分布情况。（　　）

（2）在数据预处理中，所有重复值都应该被删除。（　　）

（3）箱线图中的“须”部分表示数据中的异常值。（　　）

（4）K-Means 聚类算法的聚类结果不受初始质心位置的影响。（　　）

（5）轮廓系数值越小，聚类结果的质量越高。（　　）