

基于 RFM 的百货商场会员画像

摘要：本文利用该大型百货商场提供的附件信息，建立了该商场的会员的数据标签。对于会员的消费特征刻画，论文建立了不同的三个维度(消费层面，季节层面，生理层面)分析会员的消费特征。消费层面包括在不同价格范围内会员的消费次数，不同折扣区间会员的消费频次，会员消费总金额等统计学信息、季节层面包括不同季节对会员消费的影响和生理层面包括会员年龄信息，会员性别信息以及其对消费的影响。使得会员画像更加完整。为了比较会员与非会员的消费差异，论文从消费总金额，消费频次，单次消费金额三个方面进行比较，最终体现出会员对商场的价值。

为刻画会员的购买力，定量计算会员对于商场的价值，本文使用了传统的 RFM 模型。选取会员最后一次消费的时间间隔(R)、消费频率(F)、总金额(M)作为衡量指标，再根据层次分析法求出相应指标的权重，并计算出每个会员的得分，从而得出会员对商场价值的量化结果。利用 K-均值聚类的方法，根据会员得分进行聚类，刻画出当前会员的分类情况，从而使得商场可以更方便对会员群体进行分类管理。

为了判断会员所处的生命周期，本文利用已构建的 RFM 模型中的指标，再次使用 K-均值聚类的方法对该商场的会员进行聚类，将现有会员划分为活跃会员、一般会员、沉寂会员三类，对于不同的会员采取不同的管理方式，可以有效促进会员的价值增加。

会员入会的时间区间内，会员的状态在不断的变化。通过数据分析外面得出，激活非活跃会员所需的资金远比发展新会员要少。论文通过分析会员的相关消费数据，分析其在活跃时期与不活跃时期的消费情况，可计算出各非活跃会员激活

率，激活率越高的会员被激活的概率越大。同时，本文以会员的折扣商品消费金额在总消费金额中的占比作为会员对促销活动敏感度的反映，构建回归模型分析。结果表明，折扣活动有助于促进会员的消费。商场可对热门商品及其关联商品推出相应促销活动，同时通过广告投放、邮件推送等方式对促销活动进行推广。

关键词：会员画像；价值量化；RFM 模型；K-均值聚类；数据挖掘；营销策略关联规则；挖掘连带率；

目录

一、问题重述.....	1
二、问题分析.....	2
2.1 问题一分析.....	2
2.2 问题二分析.....	3
2.3 问题三分析.....	3
2.4 问题四分析.....	4
2.5 问题五分析.....	4
三、模型假设.....	4
四、符号说明.....	5
五、模型的建立与求解.....	6
5.1 数据预处理.....	6
5.2 问题一的模型建立与求解.....	7
5.2.1 商场会员的消费特征分析.....	7
5.2.2 会员群体与非会员群体的消费特征.....	12
5.2.3 会员群体给商场带来的价值.....	14
5.3 问题二的模型建立与求解.....	15
5.3.1 模型建立.....	15
5.3.2 模型介绍.....	15
5.4 问题三的模型建立与求解.....	22

5.5 问题四的模型建立与求解.....	23
5.5.1 因子分析模型.....	23
5.5.2 模型求解和分析.....	24
5.6 问题五的模型建立与求解.....	25
六、模型评价与优化.....	28
6.1 模型评价.....	28
6.1.1 模型的优点.....	28
6.1.2 模型的缺点.....	29
6.2 模型优化.....	29
七、参考文献.....	30
八、附录.....	30
附录 1 代码.....	30
附录 1.1 SQL 查询语句.....	30
附录 1.2 R 绘图代码.....	35
附录 1.3 SPSS 交互命令.....	40
附录 2 查重报告.....	42

一、问题重述

近几年来，随着网络的发展，“宅经济”及“懒人经济”不断凸显，电商发展迅速^①。根据调研数据显示，超过一半的人们都非常频繁的使用线上购物，这导致零售行业商场的会员不断流失，给零售运营商带来了严重的损失。会员可以给商场带来高利润，运营商应该实施一些具有针对性的营销策略来加强和会员的亲密度。会员画像可以帮助运营商加强对会员的精细化管理，定期向其推送更适合的产品和服务，与会员建立稳定的关系，促进实体零售企业更好的发展。

论文需解决以下问题：

(1) 分析该商场会员的消费特征：即确定会员群体的消费喜好和消费心理、比较会员与非会员群体的差异：主要是考虑会员和非会员消费的频率和消费金额，这些信息对商场来说是足够的。说明会员群体给商场带来的价值：根据前对两个小问题的分析，很容易得出会员群体的价值所在。

(2) 针对会员的消费情况建立能够刻画每一位会员购买力的数学模型，以便能够对每个会员的价值进行识别。这个问题的关键在于确定购买力的衡量方式，建立多个指标衡量会员的购买力，又因为购买力和会员的价值具有很高的关联性，得出会员的购买力评价模型之后，即可分析出价值。

(3) 本问题主要是分析会员的生命周期的状态的变化。会员从刚入会到变成活跃会员最终逐渐沉寂。我们可以先确定会员的状态划分指标，再确定时间窗口中会员状态的变化。

(4) 本问题考虑非活跃会员的激活率。通过分析非活跃会员的特征来进行因子分析，再与商场的促销活动建立模型，观察会员的激活率与促销活动的关联

程度。

(5) 商家策划进行促销活动。根据会员喜好和商品的连带率，分析会员的连带消费倾向，策划促销活动。

二、问题分析

2.1 问题一分析

为了分析该商场会员的消费特征，首先要对附件 3 中会员消费明细表进行筛选，因为其中包括本店会员和分店会员。之后，以附件一中的本店会员的卡号(kh)作为唯一标识符，与附件三的会员明细表中的会员卡号(kh)进行匹配，分离出该商场会员和分店会员的消费信息，然后对该商场的会员，从消费层面（包括在不同价格区间会员的消费频次，不同折扣区间会员的消费频次，会员消费总金额等统计学信息）、季节层面（包括不同季节对会员消费的影响）和生理层面（包括会员年龄信息，会员性别信息以及其对消费的影响）分析其会员的消费特征。

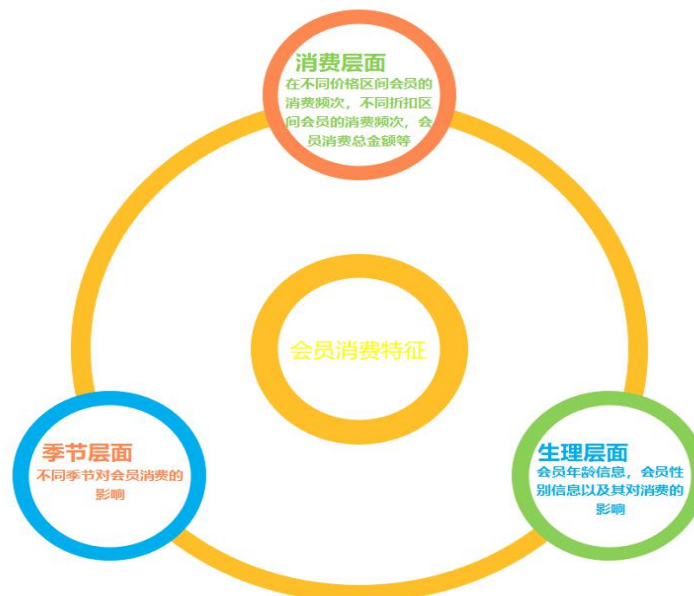


图 1.1 会员消费特征指标

对于会员与非会员群体的差异分析，我们只需要对本店的会员进行分析，因此需要将附件二中本店会员、分店会员和非会员分离，并剔除分店会员的信息。由于单据号不能作为同一笔消费的唯一标识符，我们此处选用附件三中会员消费的商品编码（spbm），消费金额（je），消费时间（dtime），收银机号（syjh）分别当作商品名称和金额（商品信息）、消费时间（时间信息）、收银机号（地点信息）的唯一标识，同时使用消费总金额，消费总频次，平均金额作为消费特征比较会员与非会员的差距。

根据前面的分析结果，结合数据可以得出会员群体给商场带来的价值。

2.2 问题二分析

对于本问题，需要针对会员的消费情况建立能够刻画每一位会员购买力的数学模型，以此来衡量每一位会员对商场的价值。为了刻画会员的购买力，我们可以使用 RFM 模型。RFM 模型是权衡客户价值和客户创利能力的重要方式^②。通过每个会员的 RFM 模型的得分来对会员进行划分，最终可得出每个会员的价值。

2.3 问题三分析

商场会员从入会到退出的过程称为会员的生命周期，但在附件中我们无法确定会员是否退会，因此在此处要重新做出会员生命周期的定义：会员的生命周期代表着会员入会后，在商场的消费状态的情况，消费状态包含活跃，非活跃，沉默等。基而消费情况的最直接体现就是会员最近一次购买商品的时间间隔天数和会员在购买商品的频率。根据问题二中提到的 RFM 模型，选取 R（会员最近一次购买商品的时间间隔天数）和 F（会员在限定时间内购买商品的频率）指标作

为聚类依据，应用 K-均值聚类法对有消费记录的会员进行状态聚类，最终可知每个会员所处的生命周期状态。

2.4 问题四分析

该问题要求计算在会员周期内（问题三中所解释的概念），求出非活跃会员的激活率。

在问题三中已经对会员进行聚类分析，分为活跃会员、一般会员和沉寂会员。从中筛选出非活跃会员，再对他们的 RFM 指标进行因子分析，计算出因子得分，并使用因子得分的方差贡献作为权重系数，对各个因子加权求和，以得出激活率。

由于商场经常举行打折促销活动，因此可以将计算出来的非会员的激活率和折扣商品总价与所有商品总价之比建立联系，使用 SPSS 软件进行相关性分析以建立模型。

2.5 问题五分析

会员的消费偏好是反映会员对不同商品的喜好程度的个性化偏好，是影响市场需求的一个重要因素，同时也是商场进行促销活动的方案的设计方向。根据会员的消费记录，使用关联数据挖掘，挖掘会员购买商品之间的内在联系。并根据相关关联规则结果策划促销活动。

三、模型假设

1. 会员从入会开始，不存在主动退会或者会员过期的情况。
2. RFM 模型中的各指标相互独立，不存在冗余的情况。

3. 附件中给定的信息全部真实准确，不存在重复的信息。

四、符号说明

表 4.1 符号说明表

符号	说明
R	最近一次消费的时间差
F	购买商品的频率
M	消费总金额
ω_i	RFM 的权重向量
ω_r	RFM 模型中指标的数量
ω_f	R 指标的权重
ω_m	M 指标的权重
a_{ij}	判断矩阵第 i 行第 j 列的元素值
x'	标准化后的值
x	标准化前的值
x^l	该指标的最大值
x^s	该指标的最小值
x_r, x_f, x_m	RFM 指标值
c_k	聚类类别
u_i	聚类中心
$J(c_k)$	距离平方和
$J(c)$	总距离平方和

符号	说明
$X = (X_1, X_2, \dots, X_p)$	可观测随机向量
$Y = (Y_1, Y_2, \dots, Y_p)$	标准化处理值
$E(Y)$	期望
A	因子载荷矩阵
$F = (F_1, F_2, \dots, F_m)$	Y 的主因子
$\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p)$	Y 的特殊因子
S_j	主因子 F_i 对 Y 的贡献率
G	因子的综合得分
f_i	各个因子
g_i	因子的贡献率

五、模型的建立与求解

5.1 数据预处理

对于众多的数据，可能会出现错误或者重复的数据，因此我们使用了 SQL 和 Excel 表格软件对附件中的数据进行预处理，处理方式包含以下方面：

- (1) 删除附件中的重复数据。
- (2) 由于我们只需要考虑本店会员的数据，因此应该根据附件一的本店会员信息，剔除附件三中非本店会员的购买信息。
- (3) 将附件二与附件三的数据连接，并分离附件三中本店会员与非会员的信息。

(4) 对于一些售价、购买数量、消费金额为负的情况，可能是由于退货的原因导致，对结果实际没有帮助，因此剔除这些信息。

(5) 处理过程中，我们发现存在 13 个购买信息的售价乘以购买数量小于购买金额的情况，显然是不合法的信息，因此要将其剔除。

5.2 问题一的模型建立与求解

首先对会员的各类消费特征进行分析，观察会员的特征，并分析会员的购物偏向，其次再对非会员的购物特征进行可视化分析，对比会员与非会员的可视化图像即能得出会员对商场消费的贡献。

5.2.1 商场会员的消费特征分析

(1) 消费层面

使用 SQL 对进行分析，统计出所有本店会员的消费信息，包括最值以及均值。

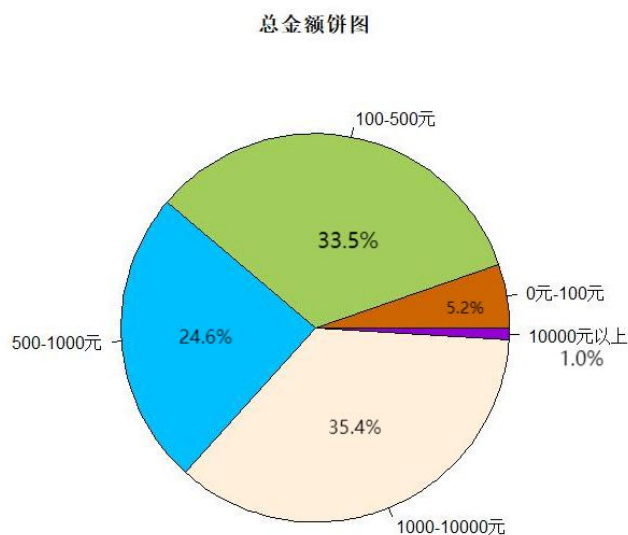


图 5.1 会员消费总金额饼图

通过图 5.1 可以得知，该商场会员群体中，消费总金额小于 100 元的人数比

例占总会员人数的 5.2%，消费总金额在[100,500]元区间的人数比例占总会员人数的 33.5%，消费总金额在[500,1000]元区间的人数比例占总会员人数的 24.6%，消费总金额在[1000,10000]元区间的人数比例占总会员人数的 35.4%，消费总金额在 10000 元以上的人数比例占总会员人数的 1.0%。可以看出，消费的金額区间基本在 100 元到 10000 元之间。

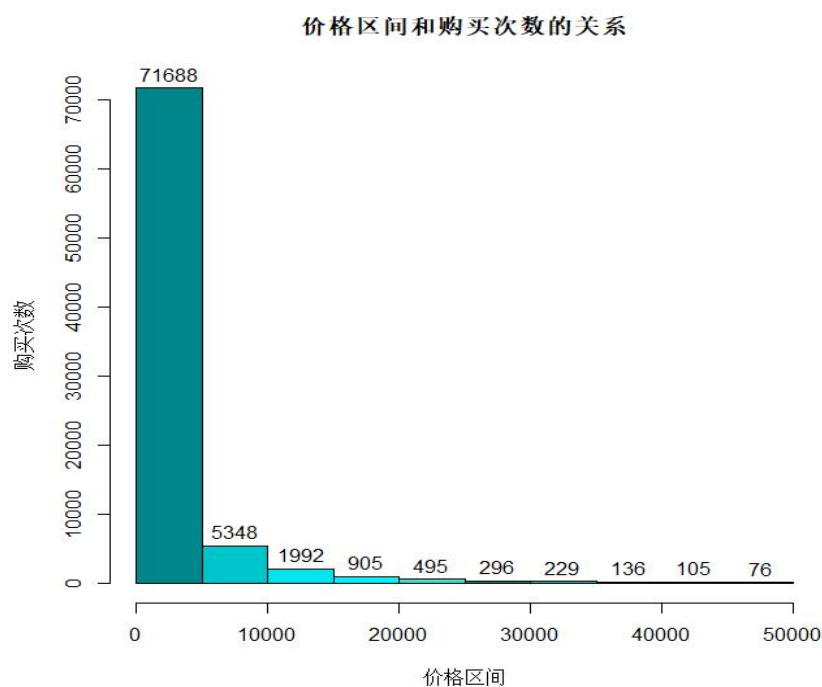


图 5.2 会员消费的价格区间与消费频次直方图

(注：价格区间仅选取 0 到 50000 元，50000 元以上因为购买次数较少使统计图不直观)

通过图 1.2 可以得知，该商场会员群体中，购买次数较多的集中在 10000 元以下，10000 到 50000 元的购买次数减少，50000 元以上由于过于稀疏在统计表中舍去。其中一件商品的最高消费金额为 1342515 元，第二最高消费金额为 283455 元，相差较大，后面的差距逐渐减小。

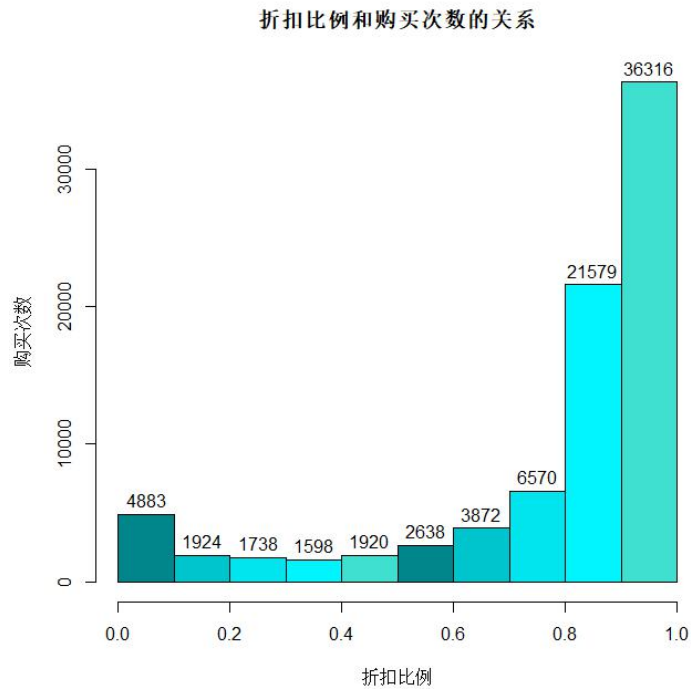


图 5.3 会员消费的折扣区间与消费频次直方图

通过图 5.3 可以得知，该商场会员群体中，会员购买商品的打折力度在两端分布较多，尤其以原价或者八到九折购买商品的会员购买数量最多，同时非常大折扣如一到两折购买的会员购买数量相较于其他折扣略多。

(1) 季节层面

根据消费时间(dtime)，从中分离出月份信息，可以得出该商场会员的消费季节偏好。选取每年的 3 月到 5 月为春季，6 月到 8 月为夏季，9 月到 11 为秋季，12 月到第二年 2 月为冬季。

通过对附件 3 的统计分析，附件 3 的会员消费明细表的数据记录时间区间为 2015 年 1 月 1 日到 2018 年 1 月 3 日，在时间维度上将其划分为 2015 年、2016 年、2017 年三年间的会员消费季节偏好情况。

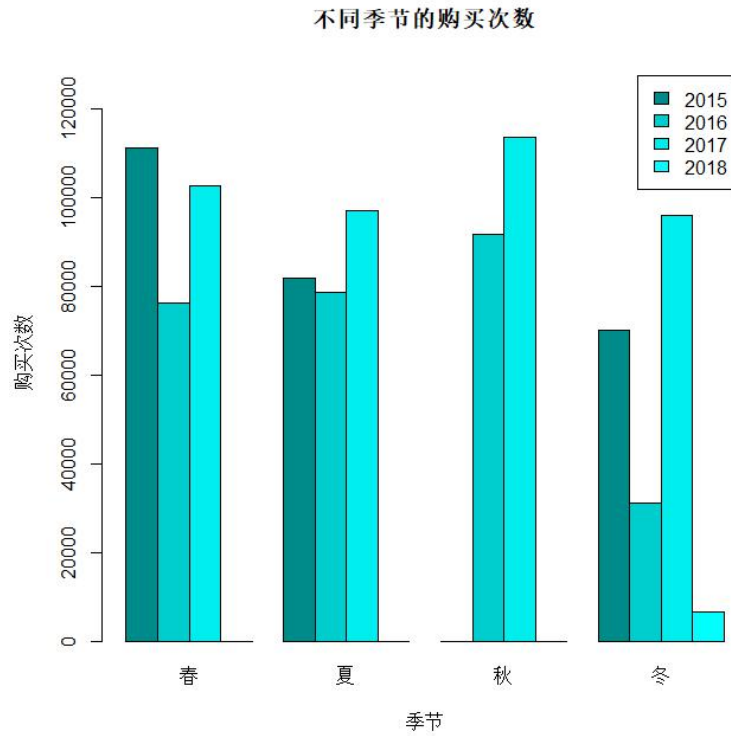


图 5.4 2015 年到 2017 年季节会员消费频次

明显可以看出，2017 年的会员消费频次较高 2016 年的会员消费频次较低（2018 年因只有冬季的一个月暂不作考虑），且 2017 年的冬季购买次数最低，2017 年秋季的购买次数最高。在季节方面，呈现春秋季购买次数较多的特点，夏季购买次数次之，冬季购买次数最少。

（2）生理层面

附件 1 该大型百货商场的会员信息表记录了该商场会员的出生年月和性别。通过会员的出生年月，统计分析出所有会员的年龄。并分别进行年龄和性别对会员消费的影响。

性别比例

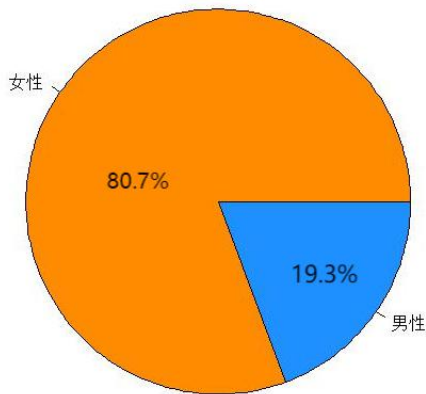


图 5.5 会员性别分布饼图

年龄比例

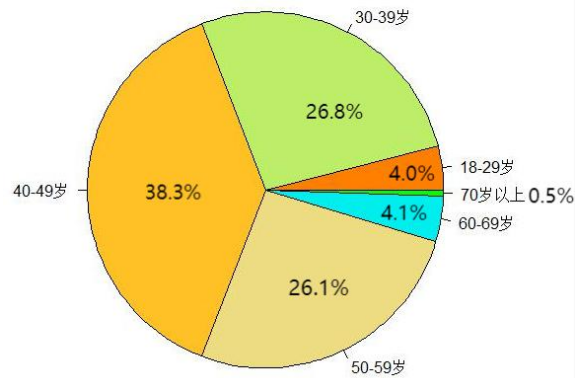


图 5.6 会员年龄分布饼图

由图 5.5 可以看出，在会员中女性占了绝大部分，为 80.7%，而男性仅为 19.3%，会员的主力为女性。由图 5.6 可以看出，会员中青年人和中年人（30-49 岁）占据了会员的绝大部分，中老年人也占据了一定比例，而相对地，青少年和老年人所占比例较少。

表 5.1 不同会员性别的消费情况

	购买总数/次	最大金额/元	最小金额/元	金额标准差/元	平均金额/元
女性	502476	1342515	0.01	3323.73	1345.53
男性	56264	200000	0.01	3680.09	1650.15

由表 1.1 可知，女性的购买次数较多于男性，但由于本身女性人数较多无法做出判断，但可以从其他方面进行分析。在单次消费中，女性拥有最大金额消费，为 1342515 元，多于男性的 200000 元，但是花费的平均金额和金额方差差异不显著。

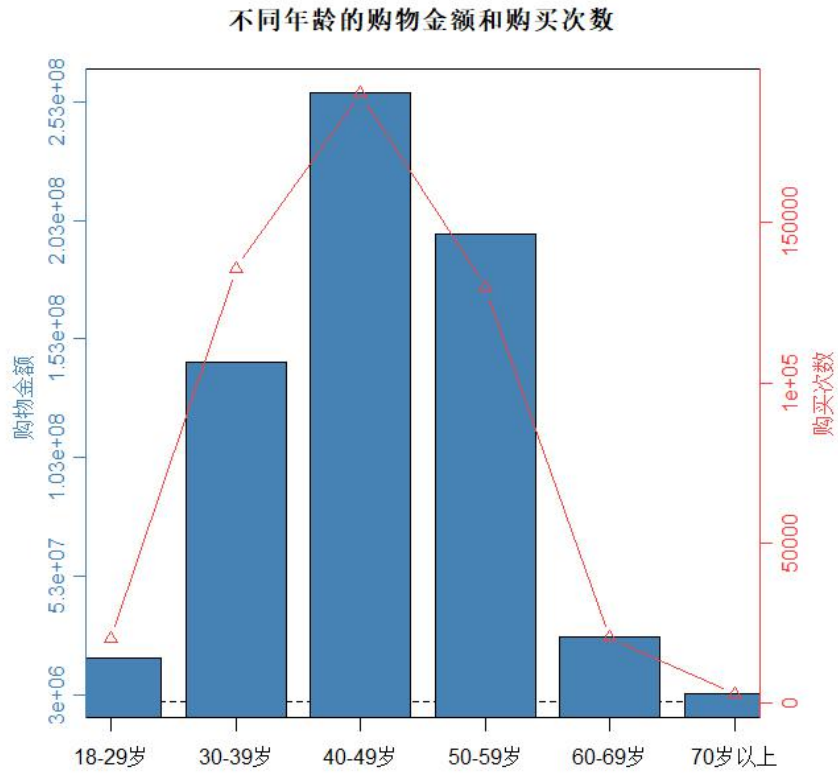


图 5.7 不同会员年龄段的消费情况

由图 5.7 可以看出，购物金额和购买次数基本呈正相关，即用户买的数量越多所对应的价格也越多。各个年龄段的购买参数也基本服从正态分布，购买金额和次数最多的集中在 40-49 岁区间内，两端的年龄购买数量和金额都较少。

5.2.2 会员群体与非会员群体的消费特征

在 5.2.1 中分析过会员群体的消费特征，在此给出非会员的消费特征。

非会员的购买金额和数量关系(10000元以下)

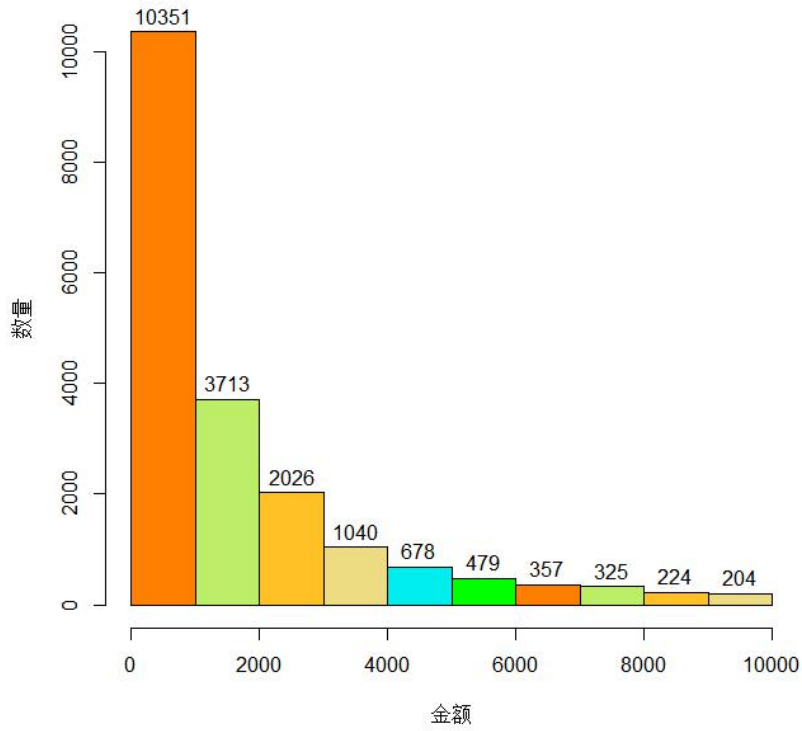


图 5.8 非会员消费的价格区间与消费频次直方图（10000 元以下）

非会员的购买金额和数量关系(10000元以上)

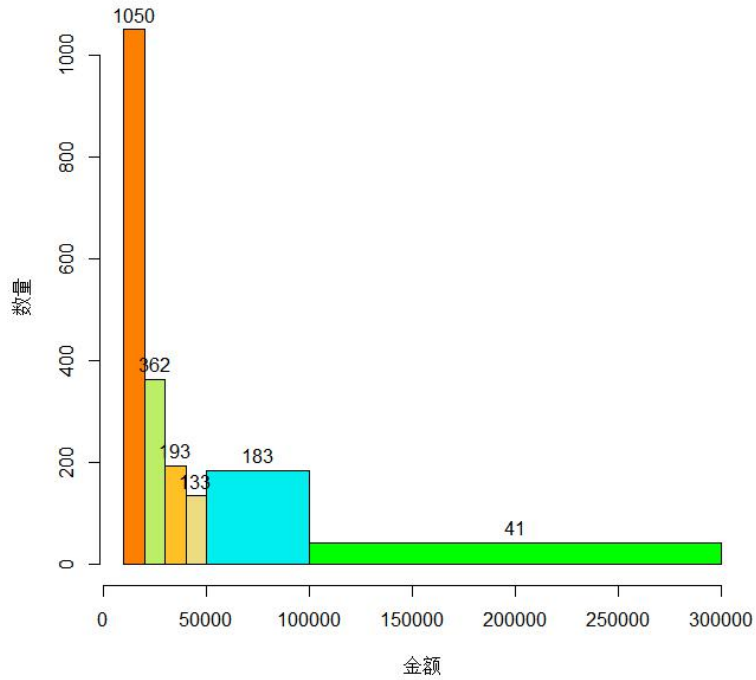


图 5.9 非会员消费的价格区间与消费频次直方图（10000 元以上）

非会员的购买折扣区间和数量关系

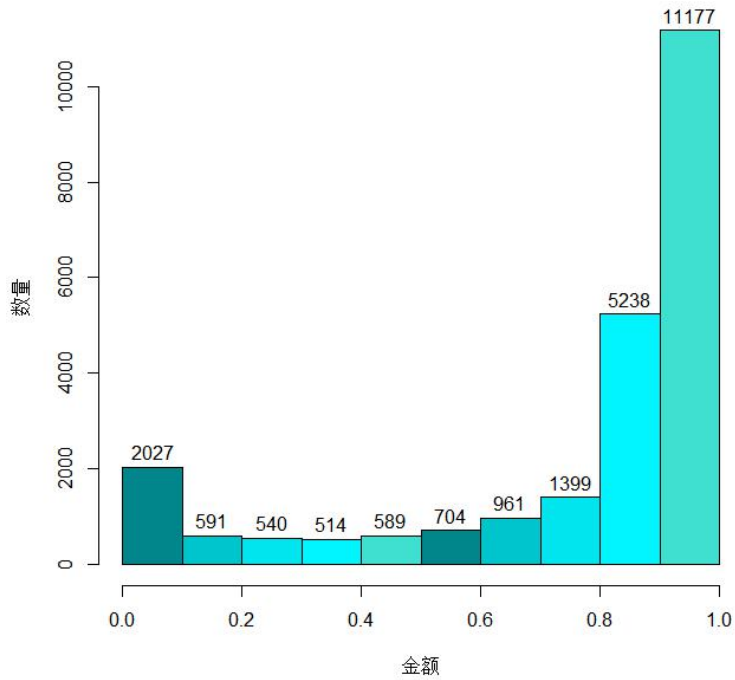


图 5.10 非会员消费的折扣区间与消费频次直方图

通过使用 SQL 进行统计分析，可以得出非会员消费群体的单个商品最大消费金额为 259244.00 元，最小消费金额为 0.01 元，平均消费金额为 4068.76 元，总金额为 86904562.43 元。根据 5.2.1 中的统计结论，可以看出会员群体的单个商品消费金额比非会员的单个商品的消费金额较多，而且对于单笔交易金额，会员群体为 135 万左右，而非会员群体仅为 30 万左右。并且，对比会员和非会员的图片看，对于购买次数来说会员群体的购买次数总体高于非会员群体。

5.2.3 会员群体给商场带来的价值

通过 5.2.1 和 5.2.2 的统计分析，明显可以得出会员的消费频率和单笔消费金额明显大于非会员，虽然非会员的数量比例更大，但对于消费总金额而言，会员群体的消费总金额是非会员的 4 倍以上。从而可以发现会员给出的价值更大。

5.3 问题二的模型建立与求解

5.3.1 模型建立

该问题的模型以 RFM 模型为基础。RFM 模型通过三项指标来描述该客户的价值状况。由于无法直接确定三个指标的权重，因此此处使用层次分析法为支撑，结合主观分析来解决指标的权重问题。最终通过 k-均值聚类算法实现会员的分类。

5.3.2 模型介绍

RFM 模型包含三个重要的指标，这些指标如下：

- 最近一次消费（Recently）
- 消费频率（Frequency）
- 消费金额（Monetary）

（1）层次分析法^③：

层次分析法(Analytic Hierarchy Process, AHP)这是一种定性和定量相结合的、系统的、层次化的分析方法^④。

算法步骤如下：

a) 建立系统的递阶层次结构模型

将决策问题分为目标层、准则层和方案层。最上层为目标层 O，即判断会员对商场的价值。中间为准则层，即 RFM 模型的三个指标。最下层为方案层，即会员价值判断的结果。

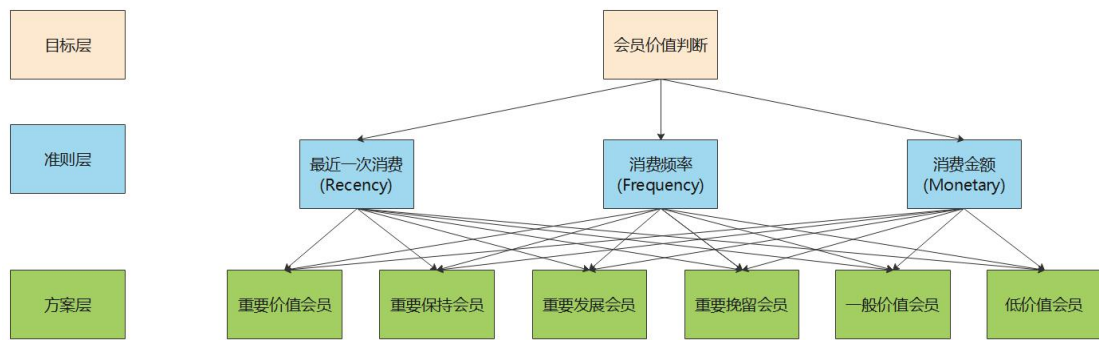


图 5.11 会员价值判断结构图

b) 构造成对比较矩阵（判断矩阵）

成对比较矩阵是表示本层所有因素针对上一层某一个因素(准则或目标)的相对重要性的比较。成对比较矩阵的元素 a_{ij} 表示的是第 i 个因素相对于第 j 个因素的比较结果，这个值使用的是 Santy 的 1-9 标度方法给出。

表 5.2 1-9 标度方法

标度	含义
1	具有同样重要性
3	一个因素比另一个因素稍微重要
5	一个因素比另一个因素明显重要
7	一个因素比另一个因素强烈重要
9	一个因素比另一个因素极端重要
2、4、6、8	上述两相邻判断的中值
倒数	因素 i 与 j 的判断 a_{ij} , 则因素 j 与 i 比较 $a_{ij}=1/a_{ij}$

接下来进行权重的判定：

■ 最近一次消费（Recently）：

主观上考虑，最近购买该商场商品的会员，是最有可能再购买东西的顾客。通过与最近消费的会员保持关系，可以提升会员的忠诚度。但由于该最近一次的消费记录和表格产生的时间也有关系，所以权重就没有那么明显。

■ 消费频率（Frequency）：

消费频率是会员对商场满意度的最直接表现，我们可以说最常购买的顾客，也是满意度最高的顾客。所以消费频率是一个很重要的指标。当时如果会员每次消费的内容都是十分便宜的商品，那么价值体现就不会有高消费的会员来的更多。

■ 消费金额（Monetary）：

消费总金额是会员价值的直接体现，在同等时间区间，消费金额最多的会员是给商场带来最高收益的会员，所以这个信息是基本上是含金量最高的指标。

表 5.3 判断矩阵 O-C

O	R	F	M
R	1	1/3	1/5
F	3	1	1/4
M	5	4	1

c) 算术平均法求权重

已知判断矩阵为

$$A = \begin{bmatrix} 1 & \frac{1}{3} & \frac{1}{5} \\ 3 & 1 & \frac{1}{4} \\ 5 & 4 & 1 \end{bmatrix}$$

那么算数平均法求得的权重向量为

$$\omega_i = \frac{1}{n} \sum_{j=1}^n \frac{a_{ij}}{\sum_{k=1}^n a_{kj}}, i = 1, 2, 3, \dots, n$$

将 RFM 各指标标准化:

由于 RFM 数据的量纲各不相同,数据的取值也存在很大的差异,因此在对各个指标进行加权之前,需要对数据进行标准化处理^⑤。

很显然,当 F、M 的值增大时,会员价值也相应增大,当 R 的值增大时,会员价值相应减小。对于指标和价值正相关的关系,使用以下公式进行标准化调整:

$$x' = \frac{x - x^s}{x^l - x^s} \quad (1)$$

对于指标和价值负相关的关系,使用以下公式进行标准化调整:

$$x' = \frac{x^l - x}{x^l - x^s} \quad (2)$$

公式(1)和(2)中, x' 为标准化后的值, x 为标准化前的值, x^l 为该指标的最大值, x^s 为该指标的最小值。

(1) K-均值聚类算法^{⑥⑦}:

使用 K-均值聚类方法来对会员群体进行聚类,分析步骤如下。

对于给定的一个包含 n 个 d 维数据点的数据集 $X = \{X_1, X_2, \dots, X_i, \dots, X_n\}$, 其中 $X_i \in R^d$, 以及要生成的数据子集的数目 K , K-均值聚类算法将数据对象组织为 K 个划分 $C = \{c_k, i = 1, 2, \dots, K\}$ 。每个划分代表一个类 c_k , 每个类 c_k 有一个类别中心 μ_i , 选取欧氏距离作为相似性和距离判断准则, 计算该类内各点到聚类中心 μ_i 的距离平方和

$$J(c_k) = \sum_{x_i \in c_k} \|X_i - \mu_k\|^2$$

聚类目标是使各类总的距离平方和 $J(C) = \sum_{k=1}^K J(c_k)$ 最小。

$$J(C) = \sum_{k=1}^K J(c_k) = \sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 = \sum_{k=1}^K \sum_{i=1}^n d_{ki} \|x_i - \mu_k\|^2$$

其中, $d_{ki} = \begin{cases} 1, & \text{若 } x_i \in c_i, \\ 0, & \text{若 } x_i \notin c_i \end{cases}$, 显然, 根据最小二乘法和拉格朗日原理, 聚类中心 μ_k 应该取为类别 c_k 类各数据点的平均值。

心 μ_k 应该取为类别 c_k 类各数据点的平均值。

K-均值聚类算法从一个初始的 K 类别划分开始, 然后将各数据点指派到各个类别中, 以减小总的距离平方和。因为 K-均值聚类算法中总的距离平方和随着类别个数 K 的增加而趋向于减小 (当 $K = n$ 时, $J(C) = 0$)。因此, 总的距离平方

和只能在某个确定的类别个数 K 下, 取得最小值。

(2) 模型的求解

a) 层次分析法计算权重:

将判断矩阵按列归一化 (每一个元素除以其所在列的和):

表 5.4 归一化矩阵

	R	F	M
R	0.1111	0.0625	0.1379
F	0.3333	0.1875	0.1724
M	0.5556	0.7500	0.6897

将归一化的矩阵按行求和:

表 5.5 归一化求和

	未归一化的权重
R	0.3115
F	0.6932
M	1.9952

将相加后得到的向量中每个元素除以 3 得到权重向量

$$W = [0.3115/3, 0.6932/3, 1.9952/3]$$

b) R、F、M 的标准化

结合公式(1), (2)和附件数据, 可得出各指标的对应值为:

表 5.6 标准化指标

<i>Rmax</i>	<i>Rmin</i>	<i>Fmax</i>	<i>Fmin</i>	<i>Mmax</i>	<i>Mmin</i>
723	100	2194	1	1981524.09	0

c) 会员聚类

会员分类体现了会员在价值上的差异,也反映了会员在行为上的变化。通过 RFM 分析可将现有顾客划分为不同的会员等级,针对不同等级的会员,采取不同的管理策略。

由图 5.7 的会员价值判断结构图,我们为聚类建立了 7 种类型,使用 SPSS 的 K-均值聚类进行分析。将客户的 RFM 平均值与总 RFM 均值比较.如果单个客户某个类别的均值大于等于总均值,则给该指标说明该指标大于平均水平,设置为 1, 否则设置为 0; 得到的结果如下:

表 5.7 各个会员的 RFM 值及分类号和标签

卡号	R	F	M	分类号	标签
8eeca369	.8491	.0091	.0056	7	0
ae6cdfbe	.9695	.0182	.0241	3	1
baa876b1	.8475	.0064	.0030	7	0
4e2b5376	.5746	.0050	.0042	7	0
af6a5229	.5169	.0005	.0005	7	0
...
dc3c2fc6	.9888	.0009	.0009	7	0
8c06e7eb	.9406	.0000	.0026	7	0

卡号	R	F	M	分类号	标签
c706369e	.9165	.0000	.0002	7	0
8ca4eace	.7897	.0000	.0006	7	0

表 5.8 每个聚类的个案数目

聚类	1	1.000
	2	27.000
	3	2053.000
	4	1.000
	5	1.000
	6	195.000
	7	31802.000
有效		34080.000
缺失		.000

d) 会员价值计算

单纯的分类型只对会员的价值做了定性的分析,这样的分析对各个类别之间的差距体现并不明显,无法体现各个类别之间定量的区别,因而对会员群体做相应的定量分析是十分必要的,根据权重的计算,我们可以定义会员的价值为:

$$V = W_r x_r' + W_f x_f' + W_m x_m'$$

其中, W_r 、 W_f 、 W_m 分别为由层次分析法分析得来的 R、F、M 指标的权重。最后,根据总得分的大小来对各类会员来进行排序。排名靠前的客户相对排名靠后的客户具有更高的价值,忠诚度更高,满意度更高对于商场来说更为重要。

表 5.9 各类会员价值指标图

卡号	$W_r x_r'$	$W_f x_f'$	$W_m x_m'$	$V \times 100$
8eeca369	.0882	.0021	.0037	9.40
ae6cdfbe	.1007	.0042	.0161	12.09
baa876b1	.0880	.0015	.0020	9.15
4e2b5376	.0597	.0012	.0028	6.36
af6a5229	.0537	.0001	.0004	5.41

卡号	Wr_{xr}'	Wf_{xf}'	Wm_{xm}'	$V \times 100$
...
dc3c2fc6	.1027	.0002	.0006	10.35
8c06e7eb	.0977	.0000	.0017	9.94
c706369e	.0952	.0000	.0002	9.53
8ca4eace	.0820	.0000	.0004	8.24

5.4 问题三的模型建立与求解

判断该商场会员处于生命周期内的状态有利于维护管理。本论文中，会员的生命周期状态分为活跃会员、一般会员、沉默会员等。

会员的消费情况会对其生命周期的状态产生影响，运用 SPSS 软件对会员的 R、F、M 三个指标数据进行 K-均值聚类，选取模型中的 R 和 F 指标作为聚类依据，建立聚类模型对会员状态进行分类。

表 5.10 会员生命周期状态划分表

类别	1	2	3
记录数	23819	15135	9803
指标 R 中心点	113.681	447.715	982.932
指标 F 中心点	16.458	5.695	4.329
指标 M 中心点	22920.953	6999.758	5456.301
定义	活跃会员	一般会员	沉默会员

从聚类中心的结果来看，类别 1 的 R 聚类中心为 113.681，F 聚类中心为 16.458，表示会员最近一次购买商品的时间距离数据截取时间的间隔天数约为 114 天，购买商品的频率约为 16 次，即会员在最近三个月内有消费且消费次数约为 16 次，相比较而言，该类别的会员足够活跃，可将此类会员划分为活跃会

员；类别 2 的 R 聚类中心为 447.715，F 聚类中心为 5.695，表示最近一次购买商品的时间距离数据截取时间的间隔天数约为 447 天，购买商品的频率约为 6 次，购买频率和时间相对于活跃会员差距很大，因此可将此类会员划分为普通会员。类别 3 的 R 聚类中心为 982.932，F 聚类中心为 4.329，表示会员最近一次购买商品的时间距离数据截取时间的间隔天数约为 983 天，购买商品的频率约为 4 次，此类会员已经基本不在该商场进行消费，可将此类会员划分为沉默会员。

5.5 问题四的模型建立与求解

5.5.1 因子分析模型

首先建立因子分析模型，根据相关性大小将变量进行分类，让同类变量之间的相关性较高，不同类变量之间的相关性较低，每一类变量代表一个公共因子。

假定 $X = (X_1, X_2, \dots, X_p)$ 为可观测随机向量，用 $Y = (Y_1, Y_2, \dots, Y_p)$ 表示标准化处理的 X ，期望 $E(Y) = 0$ 。协方差矩阵 $D(Y) = \sum F = (F_1, F_2, \dots, F_m) (m < p)$ 为不可观测的随机向量，且 $E(F) = 0$ ，协方差矩阵 $D(F) = I_m$ ，设 $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p)$ 与 F 不相关，其期望 $E(\varepsilon) = 0$ ，协方差为

$$D(\varepsilon) = \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \vdots \\ \sigma_p^2 \end{pmatrix} = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2)$$

假设随机变量能够用来表示以下模型：

$$\begin{pmatrix} Y_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \varepsilon_1 \\ Y_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + \varepsilon_2 \\ \vdots \\ Y_p = a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + \varepsilon_p \end{pmatrix}$$

将该模型称为正交因子模型，矩阵表示为 $Y = AF + \varepsilon$ 。

其中，系数 $A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{pmatrix}$ 待估计，称为因子载荷矩阵。其中

$a_{ij} (i=1,2,\dots,p, j=1,2,\dots,m)$ 为第 i 个变量在第 j 个因子上的载荷。

另外， $F = (F_1, F_2, \dots, F_m) (m < p)$ 中的各分量称为 Y 的主因子， $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p)$ 中的各分量称为 Y 的特殊因子。

将载荷矩阵中 A 中每列元素的平方和

$$S_j = \sum_{i=1}^m a_{ij}^2, j = 1, 2, \dots, m$$

称为主因子 F_i 对 Y 的贡献率，用来衡量其对于主因子的相对重要性。

5.5.2 模型求解和分析

使用 SPSS 对筛选的非活跃会员进行指标因子分析，得到因子的综合得分 G ，计算方法如下

$$G = \sum_{i=1}^3 f_i \times \left(\frac{g_i}{g_1 + g_2 + g_3} \right)$$

其中 f_i 为各个因子， g_i 为各个因子的贡献率。可以计算出流失会员和一般会员的激活率如下表所示：

表 5.11 各类会员的激活率

流失会员卡号	激活率	一般会员卡号	激活率
cf4add41	43.79%	fc4e49a8	4.50%
14ec5f13	46.98%	2777c933	19.41%
c45aa115	43.18%	20a97091	38.28%
6f12c167	49.90%	dd80bbdc	10.42%
a0993bdc	42.74%	0e06e863	25.03%
...
20eaed3d	49.48%	46c6721e	41.91%
a66e1367	40.00%	1cd34d9e	1.72%
da876aa5	43.50%	d4286f41	1.20%

建立多元现行回归分析模型来分析激活率和折扣活动的关系。假设 Y 与 X 存在以下关系：

$$Y = \beta_0 + \beta_X X + \beta_R X_R + \beta_F X_F + \beta_M X_M + \varepsilon$$

上式中 $\varepsilon \sim N(0, \sigma^2)$ 。

利用最小二乘法进行拟合，可以得知流失会员的激活率和促销活动之间的关系：

$$Y_1 = -0.62 + 4.57 \times 10^{-13} X_1 + 9.22 \times 10^{-5} X_{R1} + 8.26 \times 10^{-3} X_{F1} + 6.23 \times 10^{-6} X_{M1}$$

一般会员的激活率与促销活动的关系为：

$$Y_2 = -0.63 + 1.66 \times 10^{-2} X_2 + 9.13 \times 10^{-4} X_{R2} + 1.35 \times 10^{-2} X_{F2} + 2.24 \times 10^{-9} X_{M2}$$

对比以上两式，可以看出在两个会员群体中激活率和商场促销活动均呈正比例关系，且其对一般会员的促进效果优于其对流失会员的促进效果。

5.6 问题五的模型建立与求解

会员的消费偏好是反映会员对不同商品的喜好程度的个性化偏好，是影响市场需求的一个重要因素，同时也是商场进行促销活动的方案的设计方向。根据附

件 3 中的本店会员的消费记录，联合附件 4 中的商品信息表，统计出会员的消费偏好即会员消费数量较多的十件商品。根据会员的消费偏好，使用关联规则统计出会员一次消费中商品的连带情况，并进行统计。对消费连带率较高的商品组合进行分析，给出消费促销活动策划。

首先根据附件 3 中的会员消费记录，使用 SQL 统计出会员消费中购买数量前十的商品。

表 5.12 会员消费数量排名前十的商品相关信息

排名	商品类别	商品名称	购买总数量
1	40102	兰芝化妆品正价瓶	9863
2	40101	欧舒丹化妆品瓶	7726
3	40102	欧莱雅化妆品系列支	6400
4	40101	圣罗兰纯口红/唇釉/纯魅/蜜糖/金粹	5702
5	40101	后瓶	5624
6	70101	素然正价件	5389
7	40101	Fancl 无添加瓶	4547
8	70302	爱慕内衣正价件	3568
9	70101	玛丝菲尔淑女装系列 A 件	3318
10	40101	雅诗兰黛特润修护肌透精华露 50ml	3050

针对会员消费情况，统计会员一次消费的所有消费商品购买情况，即统计一次消费中的商品编码。在会员的消费明细表中，一个会员可能会进行多次消费并

且一次消费中购买多个商品编号相同的商品，即单单会员的卡号不能判断出消费记录是否会员的一次消费。会员的卡号仅仅提供会员的信息，加上会员消费的时间以及收银机号即可以判断出消费记录是否是同一次消费记录。因为一般不会有重复的收银机号，并且一个收银机只有一个收银机号。一个收银机号对应一个在商场中固定位置的收银机。即会员卡号、消费时间、收银机号即确定了谁在什么时候，在什么地点进行了消费。

使用 Clementine 软件对会员一次消费中购买的商品进行关联规则挖掘，结果如表 5.13 所示.

表 5.13 会员消费前十的商品的关联商品

排名	商品名称	相关关联商品 1	相关关联商品 2
1	兰芝化妆品正价瓶	Fancl 无添加瓶	迪奥梦幻美肌修颜乳
2	欧舒丹化妆品瓶	banxiaoxue A 件	雅芳婷床用系列 F 床
3	欧莱雅化妆品系列支	科颜氏高保湿面霜 125ml	梦迪亚 美易 B.5 无
4	圣罗兰纯口红/唇釉/纯魅/蜜糖/金粹	兰蔻唇膏/玫瑰唇釉	香奈儿可可小姐/邂逅 香水喷装 100ml
5	后瓶	兰蔻精准淡斑臻白精华 乳 30ml	六福黄金系列正价件
6	素然正价件	IMMOBILE A 件	雅诗兰黛鲜亮焕采洁 面乳 125ml
7	Fancl 无添加瓶	Fancl 无添加瓶	SONG OF SONGB.5 件
8	爱慕内衣正价件	香奈儿可可小姐/邂逅香 水喷装 100ml	圣罗兰纯口红/唇釉/ 纯魅/蜜糖/金粹
9	玛丝菲尔淑女女装系列 A 件	娇韵诗恒润奇肌保湿面 膜 50ml	娇韵诗恒润奇肌保湿 面膜 50ml
10	雅诗兰黛特润修护肌透精华露 50ml	阿玛尼持色迷情唇膏 4.2ml	圣罗兰气垫粉底液

根据表 5.13，可以看出关联规则挖掘的结果，即会员消费排名在前十的商品及其相关关联商品。下面基于表 x.x 给出此次促销活动的策划：

- (1) 某一品牌一次消费满 2888 元以上，享受 8.8 折优惠。
- (2) 化妆品跨品牌优惠，一次消费满 1888 元减 288 元。
- (3) 购买服饰类、床上用品，消费满 2888 元，享受 8 折优惠。
- (4) 购买珠宝首饰 18888 元以上，享受 8.8 折优惠，并凭购物小票可领取化妆品 200 元抵用券。

商场进行搭配销售的同时，商场可以调整柜台的位置，以及相关商品的位置，均以就近原则摆放。

六、模型评价与优化

6.1 模型评价

本论文首先通过数据挖掘的方法，将会员的消费特征从消费层面，季节层面和心理层面进行了分析，比较了会员与非会员的消费差距，体现了会员给商场带来的价值。从商场近几年的消费流水线和会员消费信息，通过层次分析和聚类算法，利用了传统的 RFM 模型，刻画了会员的购买力和会员价值的定量分析方式。结合会员购买力的模型，对会员生命周期做了定义，对会员的生命周期状态进行了划分。

6.1.1 模型的优点

(1) 论文使用的模型是认可度很高的 RFM 模型。并未对 RFM 模型做出任何的改动，模型足够严谨，考虑全面，数据使用 SQL、Excel、SPSS 等专业数据处理软件进行分析，并且全部进行定量的分析，可信度更高，结果将人的主观性降到很低，使得分析更加客观。

(2) 通过 K-均值聚类算法对会员进行了分类，与实际相结合，有助于商场更好的管理会员，判断会员的价值，确定该与哪些会员保持密切联系，哪些会员的关注度可以降低。

(3) 模型的拓展性和通用性很强，对于更新的数据，可以直接通过模型进行处理，基本不需要对模型进行改变。

(4) 通过数据挖掘的方式对连带消费行为进行了分析，从而可以实现更合理的推送。

6.1.2 模型的缺点

(1) 对于购买量为负的消费记录的处理可能不合理，我们其实可以将这类消费记录理解为客户的退货行为，对这些行为进行分析可以帮助商场发现退货率较高的商品，从而做出改变。

(2) RFM 模型中各指标存在一定的相关性，论文没有考虑这些指标的相关性。另外，这些指标实际上不足以概括会员的全部消费特征，比如说会员的消费喜好，单次消费金额的最值和方差等等。

(3) 对于问题五的关联规则，软件分析的结果缺乏置信度。另外，消费数量前十的商品，全是女性用品，具有关联性的商品也均是女性用品，这导致推荐策略制定时，会出现仅偏向女性的推荐的情况，导致推荐策略的偏差。

6.2 模型优化

在数据允许的情况下，在 RFM 模型中增添会员的消费喜好(S)和会员的单次消费金额的最值(H)等指标，从而可以更好的分析会员的消费特征。

深入分析模型中各指标的关联性，除去指标间的冗余信息，从而建立一个更

加准确的会员评价体系。

七、参考文献

- [1]于海洋,网络订餐平台营销模式研究[J]. 黑龙江大学学报,2016
- [2]MBA 智库-百科 RFM 模型
- [3]林盛,肖旭.基于 RFM 的电信客户市场细分方法[J].哈尔滨工业大学学报,2006,(05)
- [4]知乎 <https://zhuanlan.zhihu.com/p/38207837>
- [5]百度百科 数据标准化
- [6]张文彤,董伟.SPSS 统计分析高级教程[M].北京:高等教育出版社,2004.247-248
- [7]王千,王成,冯振元,叶金凤.K-means 聚类算法研究综述[J].电子设计工程,2012,20(07):21-24.

八、附录

附录 1 代码

附录 1.1 SQL 查询语句

```
USE shopping  
  
SELECT DISTINCT  
c3. djh, c3. dtime, c3. gzbm, c3. gzmc, c3. je, c3.  
jf,  
c3. kh, c3. sj, c3. sl, c3. spbm, c3. spmc, c3.  
syjh
```

```

INTO SELE

FROM c3, c2, c1

WHERE c3. spbm = c2. spbm and c3. je=c2. je
and c3. dtime=c2. dtime and c3. syjh = c2. syjh

and c3. kh = c1. kh and c3. sj>=0 and

c3. je>=0 and c3. sl>=0 and

c3. je<=(c3. sj*c3. sl)

```

```
USE shopping
```

```
SELECT distinct je, sum(sl) sltotal
```

```
--INTO vip1
```

```
FROM c3
```

```
GROUP BY je
```

```
HAVING je>=0
```

```
ORDER BY je
```

```
SELECT DISTINCT sj, je, sum(sl) sltotal,
je/(sj*sl) zk
```

```
INTO vip2
```

```
FROM c3
```

```
GROUP BY je, sj, sl
```

```
HAVING je>=0
```

```
ORDER BY je
```

```
SELECT DISTINCT zk, sum(sltotal) sumt
```

```
FROM vip2
```

```
GROUP BY zk
```

```
HAVING zk<=1 and zk>=0
```

```
ORDER BY zk
```

```
USE shopping
```

```
select sum(je) 总金额, avg(je) 平均金
```

额, max(je) 最大金额, min(je) 最小金

额, var(je) 金额方差,

```
sum(sl) 总数量, avg(sl) 平均数
```

量, max(sl) 最大数量, min(sl) 最小数

量, var(sl) 数量方差

```
into a1
```

```
from c3
```

```
where je>0 and sl>0
```

```
select c1. xb, sum(c3. je) 总金
```

额, avg(c3. je) 平均金额, max(c3. je) 最大金

额, min(c3. je) 最小金额,

```

sum(c3. s1) 总数量, avg(c3. s1) 平均数
量, max(c3. s1) 最大数量, min(c3. s1) 最小数
量

into a2

from c3, c1

where je>0 and s1>0 and c3. kh = c1. kh and
c1. xb is not NULL

group by c1. xb

USE shopping

SELECT kh, s1,
LEFT(dtime, 10) as mydate,
RIGHT(dtime, 8) as mytime

INTO sep

FROM c3

SELECT kh, s1,
LEFT(mydate, 2) as mymonth,
RIGHT(mydate, 4) as myyear,
mytime

INTO sep1

FROM sep

```

```

SELECT DISTINCT mymonth, myyear, sum(s1)
sltol

FROM sep1

GROUP BY mymonth, myyear

USE shopping

SELECT DISTINCT
myyear, season, sum(sltol) sltol

FROM vip3

GROUP BY myyear, season

ORDER BY myyear, season

USE shopping

SELECT DISTINCT xb, SUM(s1)
totals1, MAX(je) maxje, MIN(je)
minje, STDEV(je) varje, AVG(je) avgje

FROM c1, c3

WHERE c1. kh = c3. kh and je>0

GROUP BY xb

HAVING xb is not null

```

```
SELECT DISTINCT xb, je
FROM c1, c3
WHERE c1.kh = c3.kh and je>0 and xb is
not null
```

```
USE shopping
SELECT COUNT(je)
FROM c3
WHERE je>=0 and je<=100
```

```
SELECT COUNT(je)
FROM c3
WHERE je>100 and je<=500
```

```
SELECT COUNT(je)
FROM c3
WHERE je>500 and je<=1000
```

```
SELECT COUNT(je)
FROM c3
WHERE je>1000 and je<=10000
```

```
SELECT COUNT(je)
FROM c3
WHERE je>10000
```

```
USE shopping
SELECT xb, COUNT(xb)
FROM c1
GROUP BY xb
```

```
USE shopping
SELECT c1.kh, je, s1,
2020-RIGHT(LEFT(csn, 10), 4) as
```

```
BIRTHYEAR
INTO age
FROM c1, c3
WHERE c1.kh = c3.kh and csny is not null
```

```
AND 2020-RIGHT(LEFT(csn, 10), 4)>=18 and
je>=0 and s1>=0
```

```
ORDER BY BIRTHYEAR
```

```
SELECT BIRTHYEAR, COUNT(BIRTHYEAR)
FROM age
```

```

GROUP BY BIRTHYEAR

SELECT BIRTHYEAR, SUM(je)
totalje, SUM(s1) totals1

FROM age

GROUP BY BIRTHYEAR

USE shopping

SELECT DISTINCT c2. je, c2. s1

INTO vipje

FROM c2, c3

WHERE c2. spbm = c3. spbm and c2. je =
c3. je

and c2. dtime = c3. dtime and c2. syjh =
c3. syjh

and c2. je>=0 and c2. sj>=0 and c2. s1>=0

ORDER BY c2. je

SELECT DISTINCT c2. je, c2. s1

FROM c2

WHERE NOT EXISTS (

SELECT je, s1

```

```

FROM vipje

WHERE c2. je = vipje. je and c2. s1 =
vipje. s1

)

and c2. je>=0 and c2. s1>=0

ORDER BY je

SELECT DISTINCT c2. je/(c2. sj*c2. s1)
b1, c2. s1

INTO vipb1

FROM c2, c3

WHERE c2. spbm = c3. spbm and c2. je =
c3. je

and c2. dtime = c3. dtime and c2. syjh =
c3. syjh

and c2. je>=0 and c2. sj>=0 and c2. s1>=0

and c2. je/(c2. sj*c2. s1)>=0 and
c2. je/(c2. sj*c2. s1)<=1

ORDER BY c2. je/(c2. sj*c2. s1)

SELECT DISTINCT c2. je/(c2. sj*c2. s1)
b1, c2. s1

```

```

FROM c2

WHERE NOT EXISTS(

SELECT b1,s1

FROM vipbl

WHERE c2.je/(c2.sj*c2.s1) = vipbl.b1

```

```

and c2.s1 = vipbl.s1

)

and c2.je/(c2.sj*c2.s1)>=0 and s1>=0

and c2.je/(c2.sj*c2.s1)<=1

ORDER BY b1

```

附录 1.2 R 绘图代码

```

setwd("E:\\2020-2021-1\\数学建
模\\期末\\题目二附件")

vipcut0 <-
read.csv("vip1.csv")

#####

#####

windows() #受极端值影响

hist(vipcut0$je,vipcut0$sltotal,
al,labels =
TRUE,breaks=seq(0,1500000,50000)
,

main = "价格区间和购买次数的
关系",xlab = "价格区间",ylab = "购买
次数")

```

```

vip1 <- vector()

vip1 <- vipcut0[c(1:81270),]

#小于 50000,去除了极端值

windows()

hist(vip1$je,vip1$sltotal,lab

els =

TRUE,breaks=seq(0,50000,5000),

main = "价格区间和购买次数的
关系",xlab = "价格区间",ylab = "购买
次数",

col =

c("turquoise4","turquoise3","tur
quoise2","turquoise1","turquoise
"))

```



```

#####

#####

vip2 <- read.csv("vip2.csv")

#去除了负值和大于1的值

windows()

hist(vip2$zk,vip2$sumt,labels

= TRUE,breaks=seq(0,1,0.1),

main = "折扣比例和购买次数的

关系",xlab = "折扣比例",ylab = "购买

次数",

col =

c("turquoise4","turquoise3","tur

quoise2","turquoise1","turquoise

"))

#####

#####

vip3 <- read.csv("vip3.3.csv")

vip3 <- as.matrix(vip3)

spring <-

c(111204,76274,102518,0)

```

```

summer <-

c(81905,78687,96943,0)

autumn <- c(0,91771,113422,0)

winter <-

c(69921,31148,95796,6433)

seasons1 <-

matrix(c(spring,summer,autumn,wi

nter),4,4)

season <- c("春","夏","秋","冬

")

cols <-

c("cyan4","cyan3","cyan2","cyan"

)

windows()

barplot(seasons1,names.arg =

season,beside = TRUE,col = cols,

legend =

c("2015","2016","2017","2018"),m

ain = "不同季节的购买次数",

ylim = c(0,130000),xlab

= "季节",ylab = "购买次数")

```

```
#####

#####

f <- 500045

n <- 11695

m <- 56053

sexsl <- c(f,m,n)

sexname <- c("女性","男性","未知

")

windows()

barplot(sexsl,names.arg =

sexname,main = "不同性别的购买次数",

      ylim = c(0,500000),col =

c("coral","deepskyblue","gray75")

)

#####

#####

fje <-

c(502476,1342515,0.01,3323,1345)

mje <-

c(56264,200000,0.01,3680,1650)
```

```
sexje <-

matrix(c(fje,mje),5,2)

sexname2 <- c("女性","男性")

windows()

#barplot(sexje,names.arg =

sexname2,beside = TRUE,main = "不

同性别的购买金额参数",xlab = "性别

",ylab = "金额")

sexje <- read.csv("vip4.csv")

windows()

boxplot(je~xb,data =

sexje,main = "不同性别的消费对比

",ylim = c(0,100000),names = c("

女性","男性"),

      ylim =

c(0,1300000),xlab = "性别",ylab = "

金额",range = 10000,col =

c("brown1","cyan"))

#####

#####

jepie <-

c(47683,303155,222262,322300,909
```

```

3)

  piename <- c("0元-100元",
              "100-500元", "500-1000元",
              "1000-10000元", "10000元以上")

  piecol <- c("darkorange3", "darkolivegreen3",
             "deepskyblue", "antiquewhite1",
             "darkviolet")

  windows()

  pie(jepie, labels = piename, col = piecol, main = "总金额饼图")

#####

#####

  xbpie <- c(149641, 35685)

  xbpiname <- c("女性", "男性")

  xbcoll <- c("darkorange", "dodgerblue")

  windows()

  pie(xbpie, labels = xbpiname, col = xbcoll, main = "性别比例")

#####

```

```

#####

  age <- c(18562, 124929, 178426, 121716, 19249, 2223+352)

  agename <- c("18-29岁", "30-39岁", "40-49岁", "50-59岁", "60-69岁", "70岁以上")

  agecoll <- c("darkorange1", "darkolivegreen2", "goldenrod1", "lightgoldenrod2", "cyan2", "green")

  windows()

  pie(age, labels = agename, col = agecoll, main = "年龄比例")

#####

#####

  ageje <- c(18462316, 142898567, 256453434, 197270227, 27635877, 3284676)

  agespace <- seq(1, 6, 1)

  agename2 <- c("18-29岁", "30-39岁", "40-49岁", "50-59岁", "60-69岁")

```

```

", "70 岁以上")

  ages1 <-
c(20039, 135629, 190316, 129667, 203
27, 2756)

#install.packages("plotrix")

library(plotrix)

windows()

twoord.plot(lx = agespace, ly =
ageje, rx = agespace, ry = ages1, type
= c('bar', 'b'),

          lcol= "steelblue",

rcol = "brown1", xticklab =
agename2,

          lytickpos =
seq(3000000, 300000000, 50000000),

          main = "不同年龄的购物
金额和购买次数",

          ylab = "购物金额

", rylab = "购买次数")

#####

#####

nvip <- read.csv("nvip.csv")

```

```

nvip1 <- nvip[1:19397,]

nvip2 <- nvip[19398:21359,]

windows()

hist(nvip1$je, nvip1$sl, labels
= TRUE, breaks=seq(0, 10000, 1000),

      main = "非会员的购买金额和数
量关系(10000 元以下)",

      xlab = "金额", ylab = "数量

", col = agecol)

windows()

hist(nvip2$je, nvip2$sl, labels
= TRUE,

      breaks=c(seq(10000, 50000, 10000),
100000, 300000),

      main = "非会员的购买金额和数
量关系(10000 元以上)",

      xlab = "金额", ylab = "数量

", col = agecol)

nvipsj <-
read.csv("nvip2.csv")

```

```

windows ()

hist (nvipsj$bl, nvipsj$sl, labe

ls = TRUE,

breaks=seq(0, 1, 0.1),

main = "非会员的购买折扣区间

和数量关系",

```

```

xlab = "金额", ylab = "数量

",

col =

c("turquoise4", "turquoise3", "tur

quoise2", "turquoise1", "turquoise

"))

```

附录 1.3 SPSS 交互命令

```

GET DATA /TYPE=XLSX
/FILE='E:\2020-2021-1\数学建模\期末\
题目二附件\RFM.xlsx'
/SHEET=name 'Sheet1'
/CELLRANGE=full
/READNAMES=on
/ASSUMEDSTRWIDTH=32767.
EXECUTE.
DATASET      NAME      数 据 集
1 WINDOW=FRONT.
DESCRIPTIVES VARIABLES=R F M
/SAVE
/STATISTICS=MEAN STDDEV MIN MAX.
QUICK CLUSTER ZR ZF ZM
/MISSING=LISTWISE
/CRITERIA=CLUSTER(7) MXITER(10) CO
NVERGE(0)
/METHOD=KMEANS(NOUPDATE)
/PRINT ID(kh) INITIAL.

```

```

MEANS TABLES=R F M BY QCL_1
/CELLS=MEAN COUNT STDDEV.
IF (ZR > 0.68126 & ZF > 0.00394 & ZM>
0.00621) hyjl=ZR > 0.68126 & ZF > 0.003
94 & ZM>0.00621.
EXECUTE.
IF (ZR > 0.68126 & ZF > 0.00394 & ZM>
0.00621) Vvalue=0.3115/3 * R+0.6932/3
*F+1.9952/3*M .
EXECUTE.
COMPUTE Vvalue=0.3115/3 * R+0.6932/
3 *F+1.9952/3*M .
EXECUTE.
COMPUTE Vvalue100=Vvalue * 100.

```

EXECUTE.

模\期末\题目二附件\问题二数据.sav'

/COMPRESSED.

SAVE OUTFILE='E:\2020-2021-1\数学建

附录 2 查重报告

PaperPass旗舰版检测报告

简明打印版

比对结果(相似度):

总体: 23% (总体相似度是指本地库、互联网的综合对比结果)
本地库: 18% (本地库相似度是指论文与学术期刊、学位论文、会议论文、图书数据库的对比结果)
期刊库: 16% (期刊库相似度是指论文与学术期刊库的对比结果)
学位库: 9% (学位库相似度是指论文与学位论文库的对比结果)
会议库: 2% (会议库相似度是指论文与会议论文库的对比结果)
图书库: 5% (图书库相似度是指论文与图书库的对比结果)
互联网: 14% (互联网相似度是指论文与互联网资源的对比结果)

报告编号: 5FF04304D2DFCA4T3

检测版本: 旗舰版

论文题目: 基于RFM的百货商场会员画像

论文作者: 以撒

论文字数: 13085字符(不计空格)

段落个数: 667

句子个数: 851句

提交时间: 2021-1-21 17:55:16

比对范围: 学术期刊、学位论文、会议论文、书籍数据、互联网资源

查询真伪: <http://www.paperpass.com/check>

句子相似度分布图:



