基于 LRFMC 模型的航空公司客户 K-means 聚类分析

摘要：区分度高的客户分类的结果是企业整合营销资源、调整营销策略的重要依据。本文根据 LRFMC 模型提取了航空公司客户的部分数据，利用 K-means 聚类方法对航空公司的客户进行分类，得到出五个不同的客户群体，对客户特征进行划分总结，给出了对应营销建议。以期对该航空公司向不同价值的客户类别提供个性化服务，制定针对性的营销策略。

关键词：旅客分类；数据挖掘；LRFMC 模型；K-means 聚类；策略分析；

1 引言

随着科技的不断发展和大数据时代的来临，数据挖掘正日渐成为研究消费者行为和特征的核心技术手段之一，而聚类分析理论为数据挖掘和机器学习提供重要理论依据。在各个领域的企业决策上，聚类分析通常都被用于客户细分和刻画用户画像，以便企业精确识别客户需求，进行个性化的推送和服务。可见聚类细分是市场细分、客户细分以及消费者行为研究的有效手段。

传统旅客划分方式主要有两种：其一是根据航班的种类进行旅客区分，将客户分为商务舱旅客和普通经济旅客；其二是根据航班的累计飞行里程进行客户划分，达到一定里程标准的客户被归类为贵宾客户，其他的为普通客户。这两种划分方式过于单调，不能够很好地反映不同客户的消费特征。航空公司的促销活动和客户对航空公司的选择性、航班次数频率等因素没有考虑在内，无法满足大数据信息化的需求。近年来也有学者根据旅客特点进行分类，主要通过旅客的固有信息如性别、年龄、职业、收入等进行聚类，但在大数据时代和信息化发展的背景下，客户的特征趋向复杂化，这种方法对客户画像的描述显得不足。

因此，本文使用客户画像特征覆盖更全面的 LRFMC 模型提取航空公司的部分数据，利用 K-means 聚类方法，对航空公司的客户进行分类，划分不同类别的客户群体，对客户特征进行总结，以期对不同价值的客户类别提供个性化服务，制定针对性的营销策略提供参考。

2 文献综述

2.1 客户价值细分

胡海（2022）提出用 K-means 算法进行 RFM 模型改进，将 RFM 属性的量化值和 K-means 算法加入 RS 理论（LEM2 算法）中，最终可以有效改善相应缺陷进一步聚焦目标客户。[1] 杨佳欣（2022）通过层次分析法（AHP）计算每个属性的权重,并以传统的 RFM 模型为蓝本,构造基于 RFM 模型改进的 LFMN 模型对用户价值细分。[3] 闫春等（2020）引入客户理赔金额指标,建立 RFMC 模型，动态设置 SOM 神经网络模型的训练速度与权重向量,将模型收敛速度提高了 21.6%并实现客户细分。[4] 李为康等（2020）建立基于 RFM 模型改进的 RVMF 模型并应用层次分析法优化对老客户划分。[5]

可以看出，现有的学者对客户细分的研究大多使用 RFM 模型，或以 RFM 模型为基础并根据实际行业现状选择相对应的价值指标，并引入层次分析法等权重计算方法，均具有行业针对性。

2.2 客户聚类算法

闫春等（2020）引入客户理赔金额指标,动态设置 SOM 神经网络模型的训练速度与权重向量,将模型收敛速度提高了 21.6%。[4] 王长琼（2018）基于谱聚类算法,结合经验规则 $k \leqslant n^{1/2}$,计算 Laplace 矩阵最大特征值差,确定客户聚类数。[5] 韩世莲（2016）提出一种可以反映多种客户需求属性的模糊系统聚类方法进行客户分类，运用智能加权对动态属性进行集成，生成相应的配送策略。[7]朱沅海等（2009）在均值聚类算法中结合 PSO 算法,对总的类内离散度和进行优化,使其达到最小值,从而获取最佳客户聚类。[8] 叶苗群（2005）提出一种改进的 K-中心点聚类算法对 Web 用户的行为进行分析，用模糊相似度并模仿遗传算法中计算适应度思想，改进传统聚类算法局部最优化的特点。[9]

由于客户特征指标属性选择而不同，国内学者对客户聚类算法的选择不尽相同，都在传统的谱聚类、系统分类、均值聚类、k-中心点聚类等分类算法上进行指标的优化改良，具有强针对性。

本文拟采用 LRFMC 模型描述客户特征，该模型覆盖航空客户特点相对全面，是一种可行的用户特征提取思路。使用 K-means 聚类模型进行客户类别分析，将数据中的观测值视为具有位置和相互间距离的对象。对客户的指标计算不进行赋权和处理操作，而将不同量纲的数据指标直接进行标准化处理后直接聚类，以得到一种简便处理的可行分类结果。

3 研究内容

在大数据的背景下，用户行为和特点的分析通过检测用户群体行为来获取数据进行整体性分析。这里用户行为的信息可以是用户在交易过程中发生所有行为收集到的信息，其次也包括用户个体固有特征信息如性别年龄等。对航空公司来说，客户特征包括客户基本信息、客户消费信息、客户积分信息，如何将这些特征进行汇总归类是航空公司对客户群进行细分的重要基础。本文主要研究内容如下：

1. 数据预处理
2. 客户特征指标提取
3. 客户聚类过程
4. 类别特征总结

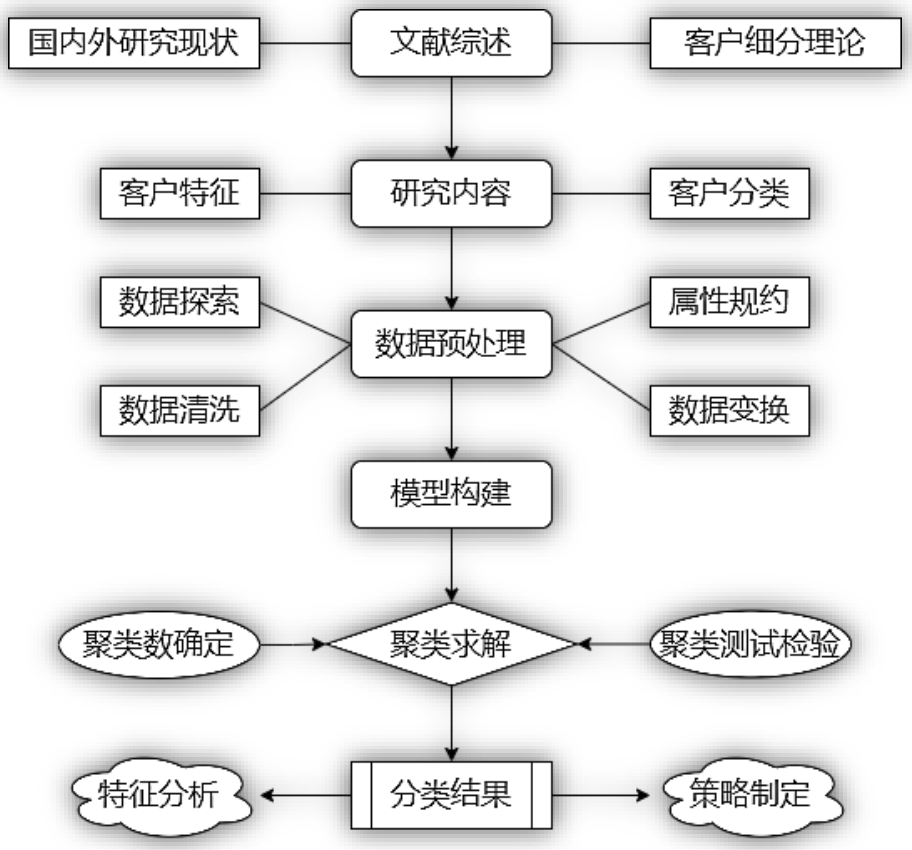
根据航空公司数据对客户进行分类，属于无监督学习，采用聚类挖掘模型，基于个体与簇之间距离的方法对客户进行类别迭代划分。在聚类模型求解之前需要向目标画像选择相应的指标，指标的选择需结合实际情况的业务进行确定，能够反映客户的关键特征。根据数据处理结果选择关键特征，进行聚类模型求解。

根据最终聚类结果，对不同的客户类别进行特征分析，比较不同类别客户的特点和价值。针对不同价值的客户制定相应的营销策略，为其提供个性化服务。

4 研究思路

用户的行为分析可以更加详细地了解用户的行为习惯，从而找出网站以及推广渠道等产品营销环境存在的问题，可以是的产品营销更加精准、有效、从而提

升企业的收益。[2] 聚类分析可以将具有相似特征的样本划分到同一簇族，对同一簇组的样本进行进一步分析，得到样本的特征分布情况。本文拟采用聚类对客户进行类别细分，将相似性程度较大地客户划分为同一类别，并根据聚类结果进行可视化分析，挖掘不同类别顾客的购买行为，以便更精确识别客户需求，采用对应的策略进行营销，为企业提供实现利益最大化的理论依据。本文研究思路如下：



图表 1 研究思路

在用户特征提取的部分采用 LRFMC 模型。

传统对消费者行为特征的分析常采用 RFM 模型。在 RFM 模型中，消费金额表示在一段时间内，客户购买该企业产品金额的总和，对现有的消费者体系具有普遍适用性，但在现有的大数据时代背景下，航空公司票价受运输距离、舱位等级等影响因素，航空公司与乘客相关时间、乘坐次数等参数指标数据具有多样性，即并不是金额越高的客户并不一定比金额低的客户价值高，具有行业的特殊性。且航空公司不定期会发放优惠券对老用户优惠票价，因此传统的 RFM 模型对现有的航空公司的数据处理是不够的（表为 RFM 模型与 LRFMC 模型对比）。

表格 1RFM 模型与 LRFMC 模型对比

模型	L	R	F	M	C
传统 RFM 模型		客户最近一次购买日**距今时间长度**	客户一段时间内**购买该企业的产品次数**	客户一段时间内**购买该企业产品金额的总和**	
航空公司 LRFMC 模型	会员**入会时间距今月数**	客户最近一次乘坐本公司飞机的时间**距今的月数**	客户一段时间内**乘坐**本公司飞机的**次数**	客户一段时间内在本公司**累计的飞行里程**	客户在一段时间内乘坐仓位所对应的**折扣系数的平均值**

基于此本文用客户在一段时间内的累计飞行里程 M 和客户在一定时间内乘坐舱位的折扣系数 C 代表消费金额。再在模型中增加客户关系长度 L，即 LRFMC 模型。

用户分类部分采用 K-means 聚类模型，将数据中的观测值视为具有位置和相互间距离的对象。它将对象划分为 K 个互斥簇，使每个簇中的对象尽可能彼此靠近，并尽可能远离其他簇中的对象。本文直接调用 pyspark.ml.clustering 库的 K-means 函数，每个簇的特性由其质心或中心点决定，距离计算欧式距离度量。由于数据量庞大，对客户的指标计算不进行赋权处理，而将不同量纲的数据指标直接进行标准化处理后直接聚类。以达到较高的聚类计算效率。

5 数据预处理

根据源数据表内容可知，源数据可能存在重复数据、缺失数据、异常数据、数据类型不匹配等干扰数据对数据分析产生干扰，因此需要对数据表进行预处理。

本节将根据源数据表，逐步进行数据的探索性分析、预处理和变换操作。探索性分析是分析源数据各个列类型、典型统计量及其分布情况，了解数据表整体情况，为后续数据选择提供参考；数据预处理包括数据清洗和属性规约，对空值、异常数据等数据进行填充和删除操作；数据变换包括属性构造和标准化，选择有价值的数据并消除量纲的影响，减小异常值的影响，为客户表的聚类分析提供可靠数据源。

5.1 数据探索分析

5.1.1 描述性统计分析

源数据表模式：

```
原数据表描述：
root
 |-- MEMBER_NO: string (nullable = true)
 |-- FFP_DATE: string (nullable = true)
 |-- FIRST_FLIGHT_DATE: string (nullable = true)
 |-- GENDER: string (nullable = true)
 |-- FFP_TIER: string (nullable = true)
 |-- WORK_CITY: string (nullable = true)
 |-- WORK_PROVINCE: string (nullable = true)
 |-- WORK_COUNTRY: string (nullable = true)
 |-- AGE: string (nullable = true)
```

图表 2 源数据表 schema（部分）

5.1.1.1基本统计量

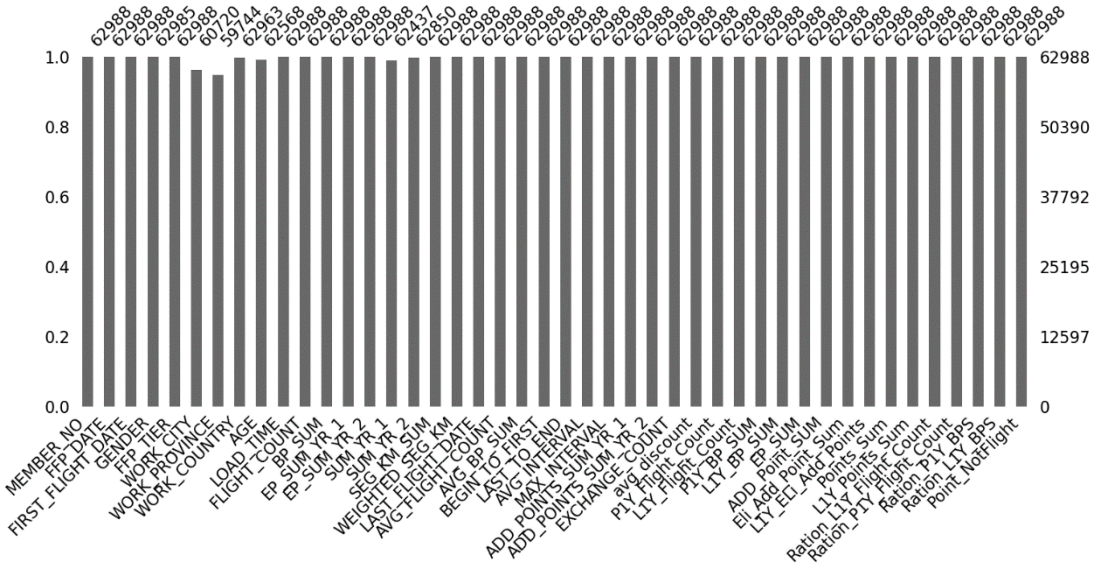
源数据表的基本统计量描述结果表 des_air_data.csv：

summary	MEMBER_NO	FFP_DATE	FIRST_FLIGHT_DATE	GENDER	FFP_TIER	WORK_CITY	WORK_PROVINCE	WORK_COUNTRY	AGE	LOAD_TIM
count	62988	62988	62988	62985	62988	60720	59744	62963	62568	6298
mean	31494.5	null	null	null	4.102162316631739	1.247327488909091E9	4190776.1304347827	null	42.47634573583941	nul
stddev	18183.21371485247	null	null	null	0.37385597940639087	4.136888094456703 7E9	2.0031436484830018E7	null	9.885914823660341	nul
min	1	2004-11-1	1905-12-31	女	4	#NAME?	#NAME?	AA	110	2014-3-3
max	9999	2013-3-9	2015-5-30	男	6	TW	TW	芬	92	2014-3-31

图表 3 源数据表描述性统计（部分）

5.1.1.2空值统计

利用 missingno 库进行缺失值可视化：



图表 4 缺失值可视化示意图

结合源数据文件统计分析可知，GENDER、AGE、WORK_CITY、WORK_PROVINCE、WORK_COUNTRY、SUM_YR_1、SUM_YR_2 列均存在空值，其中 WORK_PROVINCE、WORK_COUNTRY 缺失值比例在图像中表现较为明显，下面继续计算这些列的所含空值数和空值比例。

5.1.1.2.1 空值数

空值数：

GENDER	with null values: 3
AGE	with null values: 420
WORK_CITY	with null values: 2268
WORK_PROVINCE	with null values: 3244
WORK_COUNTRY	with null values: 25
SUM_YR_1	with null values: 551
SUM_YR_2	with null values: 138

5.1.1.2.2 空值比例

列空值比例：

GENDER	4.7628119641836543e-05
WORK_CITY	0.03600685844922842
WORK_PROVINCE	0.05150187337270591

WORK_COUNTRY	0.0003969009970153045
AGE	0.006667936749857116
SUM_YR_1	0.008747697974217311
SUM_YR_2	0.002190893503524481

5.1.2 信息分布分析

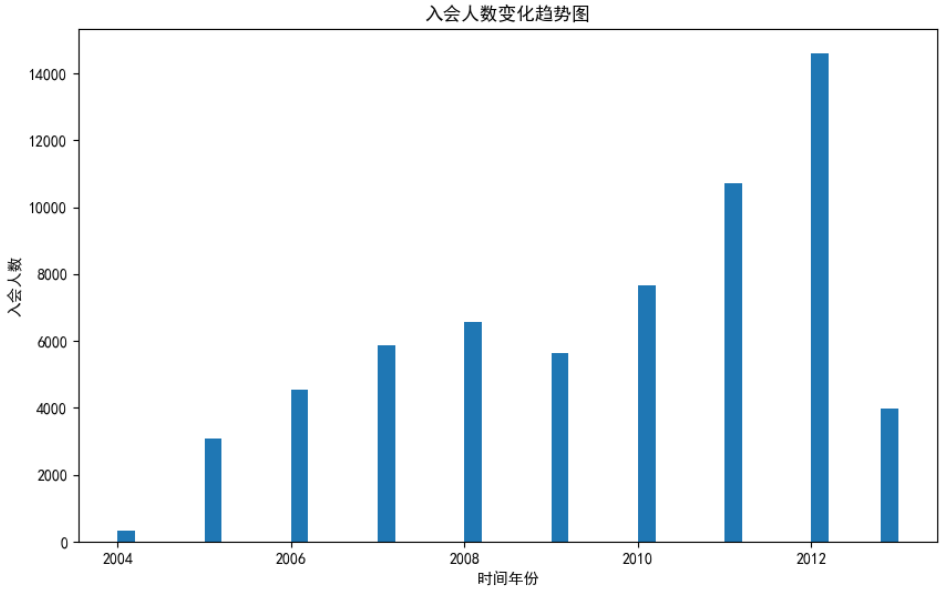
根据源数据表的描述性统计数据可知，原始数据中存在票价为空值、票价为0、折扣率最小值为0、飞行公里数大于0的记录。

结合实际航空公司的现实情况、客户买票手续和相关机票优惠业务可知，票价为空值的原因可能是乘客不存在登机记录，其他数据可能是乘客乘坐 0 折机票或积分兑换造成。

5.1.2.1客户基本信息分布分析

航空公司客户的基本信息包括客户个人信息、客户与航空公司消费手续相关信息等，本节针对客户基本信息中的入会时间、性别、会员卡级别和年龄列数据进行分析。

A. 入会人数随年份变化：

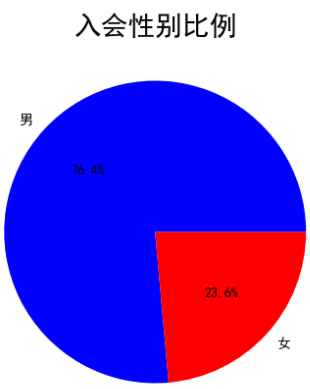


图表 5 入会人数柱形图

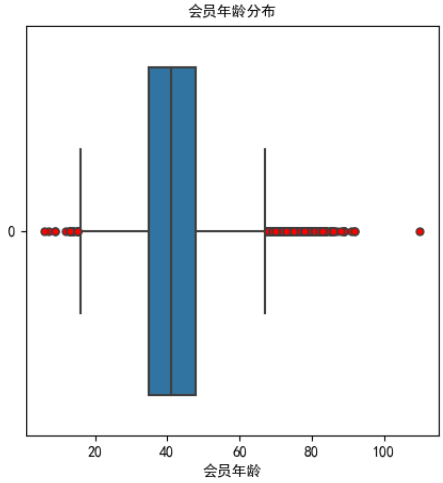
结合条形图可知，航空公司入会人数在 2004 至 2012 年间大致是随着年份

增加，并在 2012 年达到峰值，到 2013 年入会人数下降。

B. 入会会员个人基本信息（性别比例、年龄分布）：



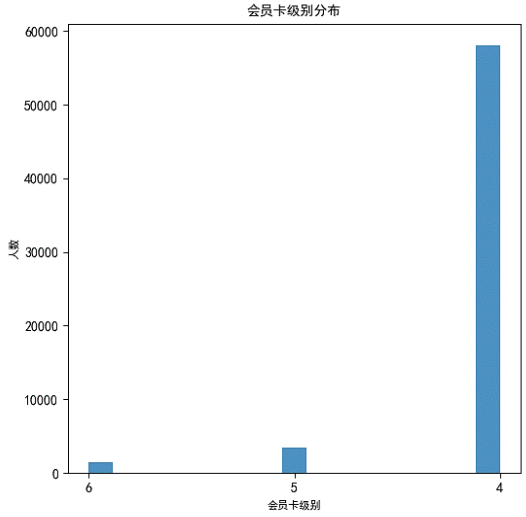
图表 6 会员性别比例饼图



图表 7 会员年龄分布箱型图

结合会员性别比例饼图和会员年龄分布箱型图可知该航空公司会员绝大部分为男性（76.4%），且会员年龄大部分在 35~50 岁，但存在一个大于 100 的异常数据。

C. 会员等级分布：



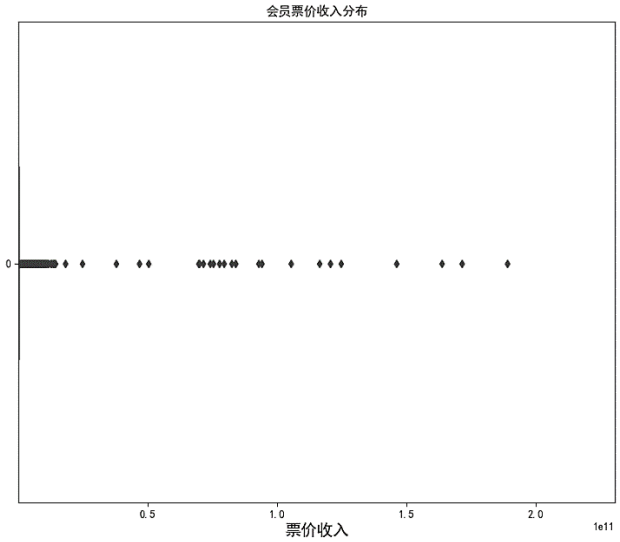
图表 8 客户会员卡级别分布条形图

结合条形图可知该航空公司的绝大多数会员为 4 级会员，接近 60000 人，5 级、6 级会员人数较少，不足 5000 人。

5.1.2.2消费信息分析

客户消费信息包括乘客在航空公司购票产生的费用，消费次数，消费内容，消费时间间隔等，本节针对客户乘机信息中的观测窗口内的飞行次数，观测窗口内的总飞行公里数，观测窗口内的票价收入，平均乘机时间间隔列数据进行分析。

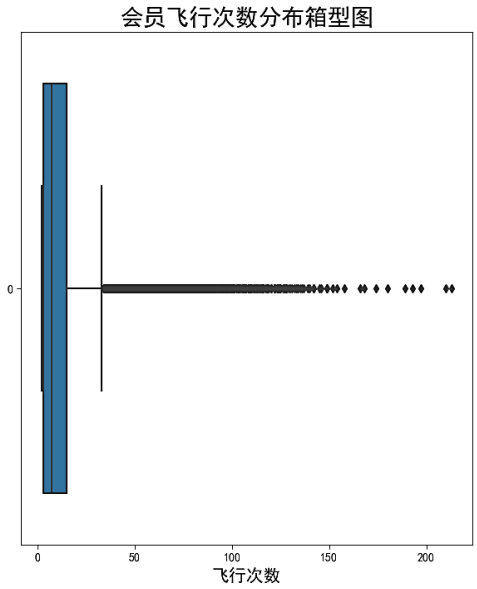
A. 会员票价收入分布：

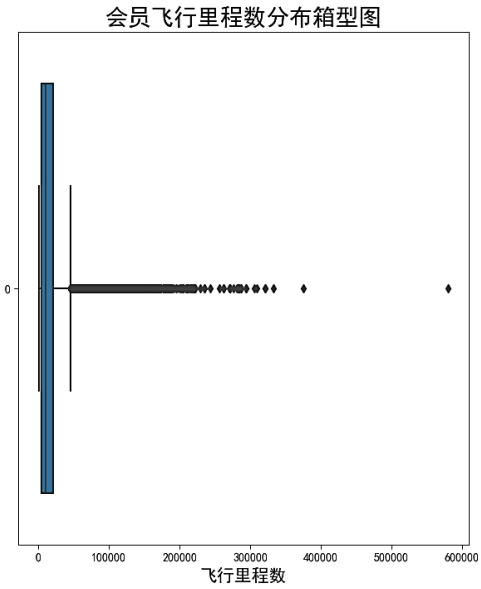


图表 9 会员票价收入分布图

根据会员票价收入可以发现，会员票价收入分布是离散的点，会员客户群体消费区间较为分散，客户群体类型多。

B. 会员飞行信息：

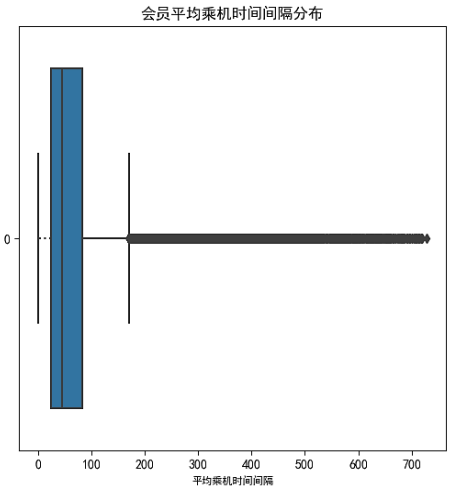


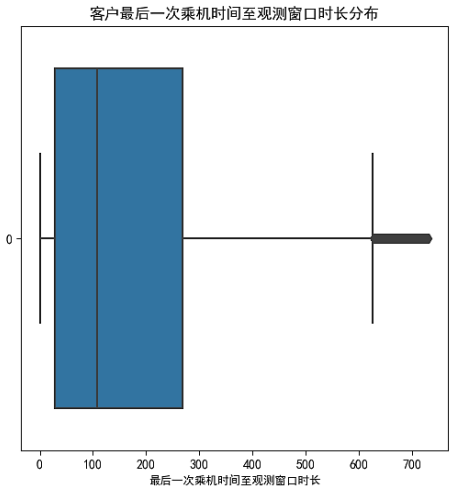


图表 10 会员飞行信息图

会员飞行信息能够代表航空公司客户的质量，飞行里程次数越多、越长，客户对航空公司的价值就越高。观测窗口内的飞行次数与观测窗口内的总飞行公里数，通过图像可以很清晰的发现：客户的飞行次数与总飞行里程数明显分为两个群体，大部分客户集中在箱型图中的箱体中，少数客户位于箱体上方，这部分客户很可能就是高价值客户。

C. 平均乘机时间间隔统计





图表 11 平均乘机时间间隔

图表 12 距最后一次乘机时间

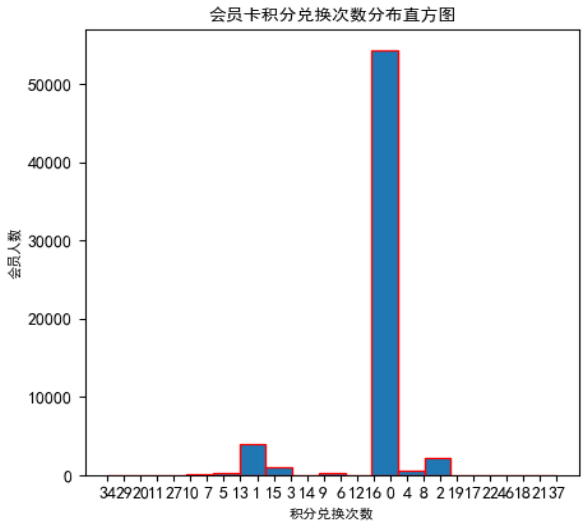
一般来说，最后一次乘机时间至观测窗口时长越短，该客户对这个航空公司的用户粘性就越大，时间间隔越短同时也说明该客户可能经常乘坐本航空公司的飞机，

是高价值客户。另外，乘机时间和间隔还可以窥探公司发展问题，如果时间间隔短的客户越来越少，说明该企业的服务或经营可能偏离时代发展，需要及时根据实际情况调整营销策略。

5.1.2.3客户积分信息

源数据中与客户积分相关的信息有积分兑换次数、总累计积分次数，本节针对客户积分信息中的积分兑换次数、总累计积分列数据进行分析。

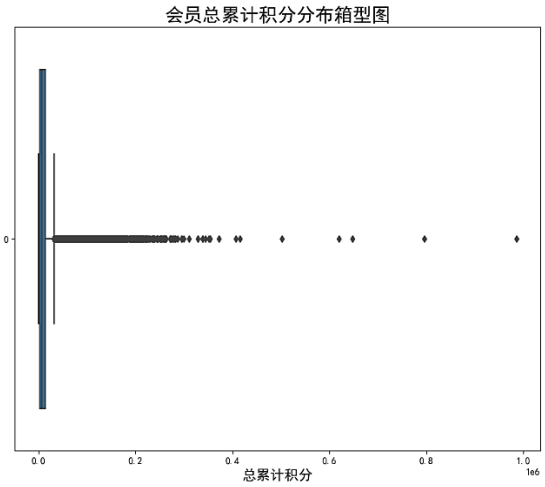
A. 不同积分兑换次数人数：



图表 13 会员卡积分兑换次数直方图

根据最高的条形柱，绝大多数兑换次数位于 0~16 次之间，这表明大部分客户很少进行积分兑换。

B. 总累计积分分布



图表 14 会员累计积分分布箱型图

大部分客户累计积分位于 400000 以内，只有个别少数客户飞行里程接近 1000000 公里。

5.2 数据预处理

本节针对航空客户数据从数据清洗、属性归纳与数据变换切入进行数据预处理。

5.2.1 数据清洗

通过数据的探索分析发现数据中存在票价为空值、票价为 0、折扣率最小值为 0、飞行公里数大于 0 的记录。由于这块的数据所占比重较小（GENDER 缺失率 4.76%，其他属性缺失率均小于 0.052%），故直接删除。

5.2.1.1异常值处理

```
# 删除年龄中的异常值
data_remove_agemg = dfread.where('AGE<100')
# 去除票价为空的记录
colsnull = ['SUM_YR_1','SUM_YR_2']
data_not_null = data_remove_agemg.dropna(how='any',subset=colsnull)
```

仅保留票价不为 0，或折扣率和总飞行公里数同时为 0 的记录

该条件通过调用 sql 语句执行，先取得源数据，再将复杂的处理和计算交给 Spark

定义一个给定名字 data_not_null 的临时表, 使用 SQL 进行查询：

```
sql_remove = "select * from data_not_null WHERE (avg_discount!=0 AND avg_discount>0)"
data_remove_diff = spark.sql(sql_remove)
```

5.2.1.2缺失值处理

根据源数据描述，发现以下数据存在缺失值：

1) 4 个类别型数据：WORK_CITY，WORK_PROVINCE ，WORK_COUNTRY ，GENDER；
2) 1 个连续型数据：AGE 有缺失值；

缺失值处理

由 5.2 中会员年龄分布箱型图可知，航空公司客户年龄较为集中，因此这里可以采用均值填充：

```
age_mean_frame = data_remove_diff.select(F.avg(data_remove_diff['AGE']))
age_mean = age_mean_frame.columns[0]
data_fillna = data_remove_diff.fillna(age_mean,'AGE')
```

5.2.2 属性规约

结合实际的航空公司业务，原 RFM 模型中 M 特征用一定时间内累计的飞行里程 M 与客户在一定时间内乘坐舱位对应的平均折扣率 C 来代替。同时考虑到会员的入会时间在一定程度上能够影响客户的价值，因此在模型中增加客户关系长度 L，作为区分客户的一种特征。本模型将以下 5 个特征作为识别客户价值的特征，即为 LRFMC 模型。

根据 LRFMC 理论， 航空公司的客户价值模型为 LRFMC，选择与其相关的六个属性，删除不相关、弱相关或冗余的属性。即入会时间、距最近乘机月数、飞行次数，飞行里程，平均折扣。原始数据中与 LRFMC 指标相关的 6 个属性为 FFP_DATE、 LOAD_TIME、 FLIGHT_COUNT、 avg_discount、 SEG_KM_SUM、

LAST_TO_END

在上述 6 个属性的基础上提取出 LRFMC 指标：

表格 2LRFMC 指标

指标	解释
L	入会时间长度
R	最近一次航班距今时间
F	航班次数
M	航班里程
C	航班折扣率

LRFMC 指标计算：

L = LOAD_TIME - FFP_DATE （属性需转换为时间格式）

R = LAST_TO_END

F = FLIGHT_COUNT

M = SEG_KM_SUM

C = avg_discount

图为 LRFMC 指标属性计算结果描述：

summary	R	F	M	C	L
count	3586	3586	3586	3586	3586
mean	220.5356943669827	13.284718349135527	20381.124093697712	1.1980257314872422	1625.0780814277746
stddev	218.34898296748423	17.651613037781388	28363.73873114869	0.17171902049199517	890.0228840925521
min	1.0	2.0	368.0	1.0	365.0
max	730.0	213.0	293678.0	1.5	3437.0

图表 15 LRFMC 指标

5.3 数据变换

5.3.1 属性构造

根据上节提取出的六个 LRFMC 指标，构造客户属性如下：

- L = LOAD_TIME - FFP_DATE

会员入会时间距观测窗口结束的月数 = 观测窗口的结束时间 - 入会时间；[单位：月]

- R = LAST_TO_END

客户最近一次乘坐公司飞机距观测窗口结束的月数 = 最后一次乘机时间至观察窗口末端时长；

[单位：月]

- F = FLIGHT_COUNT

客户在观测窗口内乘坐公司飞机的次数 = 观测窗口的飞行次数；

[单位：次]

- M = SEG_KM_SUM

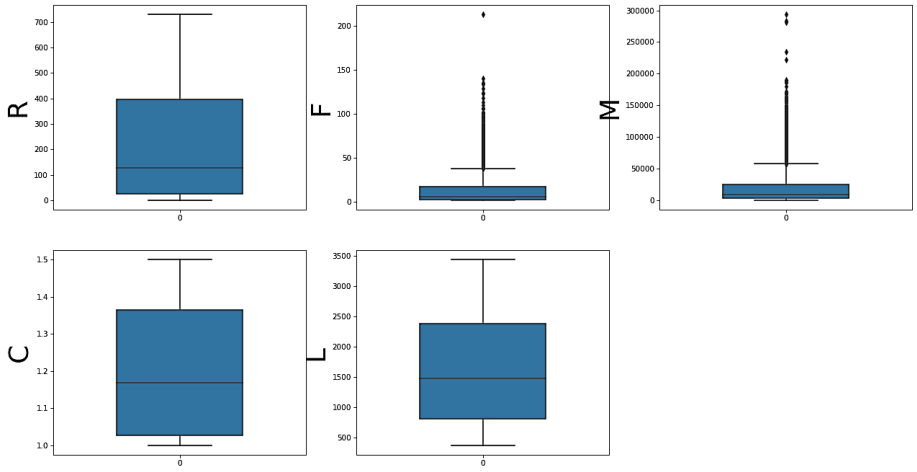
客户在观测时间内在公司累计的飞行里程 = 观测窗口总飞行公里数；

[单位：公里]

- C = AVG_DISCOUNT

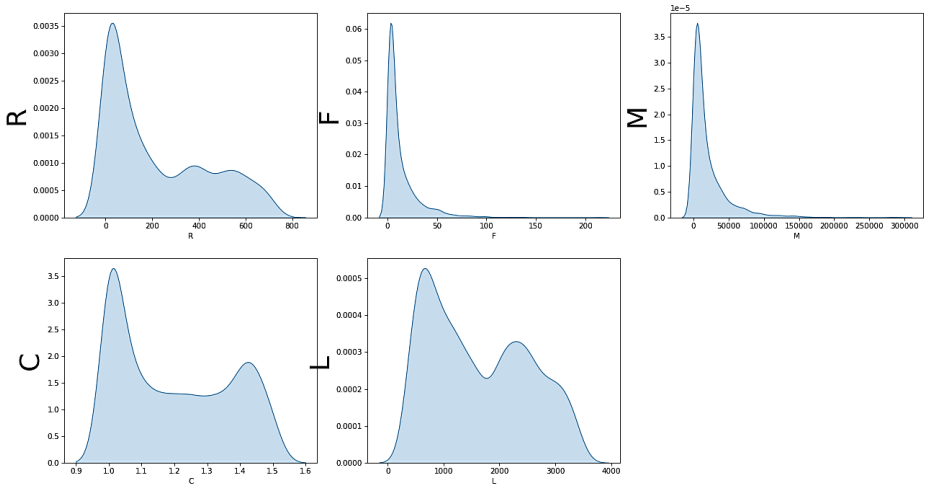
客户在观测时间内乘坐舱位所对应的折扣系数的平均值 = 平均折扣率；

[单位：无]



图表 16 **LRFMC 特征箱线图**

根据箱线图可知 F 和 M 特征存在较多异常值，说明这两个特征在客户群中分布较为分散，可以用于区分客户。



图表 17LRFMC 特征密度图

根据特征密度图可知五个属性均有较为集中的分布，且都具有明显峰值，因此五个属性可以对客户进行有价值的分类。

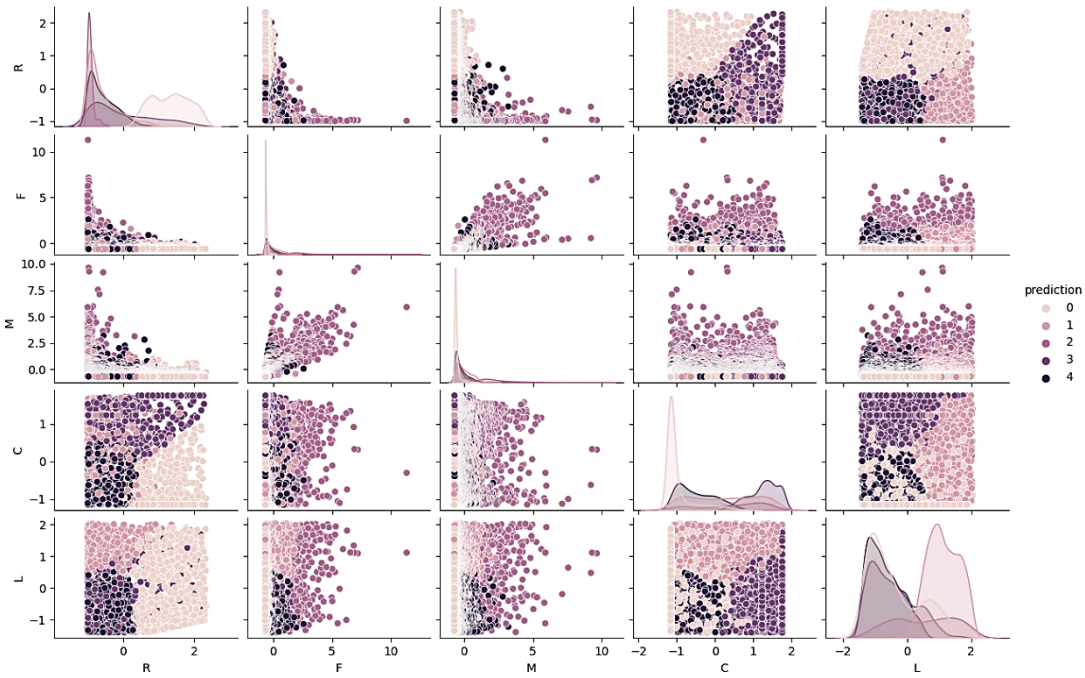
5.3.2 相关性分析

协方差值的大小并不能很好地度量两个随机变量的关联程度，值的大小受到两个变量量纲的影响，不适合用于比较。为了更好的度量两个随机变量的相关程度，引入了 Pearson 相关系数，其在协方差的基础上除以了两个随机变量的标准差，消除了量纲的影响。因此对本文数据中的相关性计算依赖 scipy 库，调用 corr 方法，使用皮尔逊相关系数计算相关系数，分别分析选择的 LRFMC 五个属性的相关性。

皮尔逊相关系数：

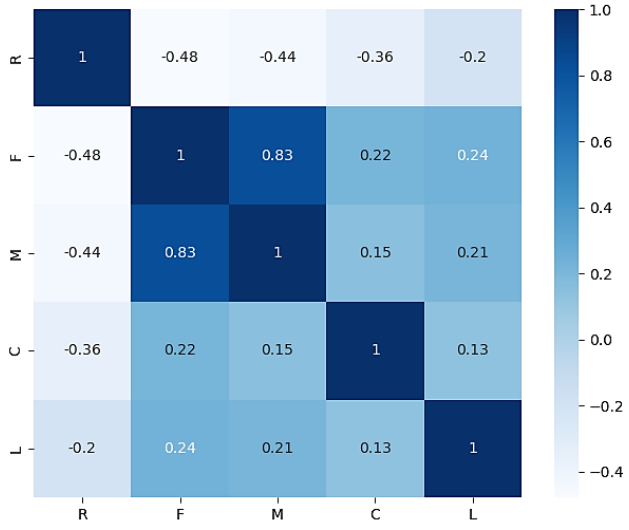
$$r_{x,y} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum_{i=1}^{n} X-\bar{X}^2} \cdot \sqrt{\sum_{i=1}^{n} Y-\bar{Y}^2}}$$

由分析得到的 pairplot 图（见图表 19）可知得到 F、M 两个特征具有相关线性关系；



图表 18 LRFMC 特征相关性 pairplot 图

由热力图（见图表 21）进一步探索特征之间相关性可知，F、M 特征相关性达到 0.83(>0.8)，具有非常强的线性关系。



图表 19LRFMC 特征热力图

5.3.3 数据标准化

聚类模型是基于距离的算法。本次分析数据集中各属性量纲不同，如数据极

值差别过大，若不处理会影响挖掘分析的效果，本例采用的数据变换的方法是标准差标准化。

采用 StandardScaler（列缩放、平移）方法标准化数据，保证每个维度数据方差为 1.均值为 0。使得据测结果不会被某些维度过大的特征值而主导。处理的对象是每一列特征：将每一维的特征列向量标准化为样本标准差为 1 或平均值为 0。

本次使用聚类前求样本标准差的计算公式为：

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}（x_i - \bar{x}）^2}$$

$\bar{x}$为x_i均值，x^*为标准化后的数据；

$$x^* = \frac{x_i - \mu}{\sigma}$$

scaler = StandardScaler()； scaler.fit(plt_pd)； data = scaler.transform(plt_pd)

```
data[:5]=====标准化numpy
[[-0.9780924   7.1796813   9.636776    0.3161926   1.0921713 ]
 [-0.95977056  6.8963814   9.285362    0.3299441   1.1123983 ]
 [-0.5658507   0.5504673   9.201582   -0.62410754  0.47412348]
 [-0.9964142   5.933162    5.611216    1.1669291   1.2236469 ]
 [-0.71242553  0.4938074   7.5578814  -1.0014334   0.63931084]]
```

图表 20 客户数据标准化

6 建模过程

6.1 模型准备

6.1.1 符号说明

A. 无标签数据集：

$$X = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \cdots \\ x^{(m)} \end{bmatrix}$$

B. 最小化损失函数：

$$E = \sum_{i=1}^{k}\sum_{x \in C_i} \|x - \mu_i\|^2$$

C. 其中μ_i为簇C_i的中心点：

$$\mu_i = \frac{1}{|C_i|}\sum_{x \epsilon C_i} x$$

6.2 建立模型

K-means 算法是一种得到最广泛使用的基于划分的无监督分类算法，把 n 个对象分为 k 个簇，以使簇内具有较高的相似度。相似度的计算根据一个簇中对象的平均值来进行。K-means 算法采用迭代更新的思想，首先随机地选择 k 个对象，每个对象代表一个簇的初始均值或中心。对剩下的每个对象，根据其与各个簇中心的欧式距离，将它分配到最相似的簇。然后 K-means 算法迭代地改善簇内变差。对于每个簇，它使用上次迭代分配到该簇的对象，计算新的均值。然后使用更新后的均值作为新的簇中心，重新分配所有对象。迭代继续，直到分配稳定，即本轮形成的簇与前一轮形成的簇相同。

一、 具体步骤：

1. 在样本中随机选取 k 个样本点充当各个簇的中心点$\{\mu_1，\ \mu_2, \dots, \mu_k\}$；
2. 计算所有样本点与各个簇之间的距离$dist$（$x^{(i)}, \mu_j$）,然后把样本点划入最近的簇中;

$$dist\left(x^{(i)}, \mu_j\right) = \sqrt{\sum_{t=1}^{n}(x^{(i)}{}_t - \mu_{jt})^2}$$

$$x^{(i)} \epsilon \mu_{nearset}$$

3. 根据簇中已有样本点重新计算簇中心；

$$\mu_i = \frac{1}{|C_i|}\sum_{x \epsilon C_i} x$$

4. 重复步骤 2、3；

二、 模型处理过程：

输入:聚类个数 k；包含 n 个对象的数据集 X={x1, x2, …, xn};

输出:k 个簇$\{S_1,\ S_2, \dots,\ S_k,\}$。

6.3 模型求解

6.3.1 实验环境

Pycharm+MiniConda+pyspark（spark-3.3.1，python3.9）；

可视化：seaborn；pyplot；missingno；

数据处理：pyspark.sql；pyspark.ml；sklearn；pandas；numpy。

6.3.2 客户聚类

6.3.2.1聚类数确定

本文使用手肘法确定聚类数，由 SSE（误差平方和）确定指标：

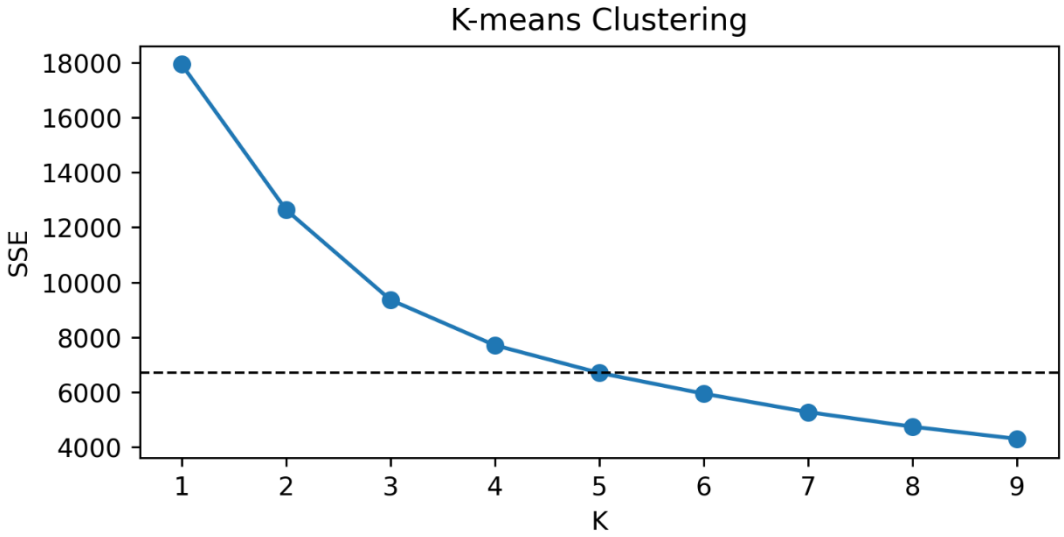
$$SSE = \sum_{i=1}^{k} \sum_{p \epsilon C_i} |p - m_i|^2$$

其中C_i为第 i 个簇，m_i为第 i 个质心，p 为属于C_i的数据点。

使用 k-means 聚类对客户进行分类，分析 k={2，3，...，9}时的误差平方和，选择最优聚类个数：

```
for k in range(1,10):
    model = KMeans(n_clusters=k, random_state=123, n_init=20)
    model.fit(airline_scale)
    sse.append(model.inertia)
```

当聚类个数未达到最优个数 k 时，随着聚类个数的增加，SSE 下降较快。达到最优个数以后，SSE 下降缓慢。可视化图形斜率变化最大处即为最优 k 值，由折线图（图表 21）可知 k=5 时下降缓慢，可以取 k 为 5。



图表 22 不同聚类数对应 SSE

6.3.2.2 KMeans 聚类训练

采用 KMeans 聚类算法对客户数据进行客户分群，由 SSE 确定聚类数 k=5，fit 用于计算训练数据的均值和方差，转换训练数据，Transform 将训练数据转换成标准正态分布。

这里聚类求解直接调用 pyspark.ml.clustering 库中的 KMeans 方法：

```
kMeans = KMeans(k=5,seed=1)
model = kMeans.fit(data_ana)
predic = model.transform(data_ana)
```

得到特征向量分类表：

features	prediction
[-0.9779559373855...	2
[-0.9596366882324...	2
[-0.5657717585563...	2
[-0.9962752461433...	2
[-0.7123261690139...	2

图表 23 求解：特征向量分类表

聚类中心：

[array([1.32175004, -0.57821878, -0.55743528, -0.96067712, -0.3459337]),
array([-0.58440156, 0.03181631, -0.00528305, 0.23449001, 1.12034427]),
array([-0.91620134, 2.36309184, 2.39013924, 0.43661927,
0.48336828]),
array([-0.01395498, -0.2803988 , -0.31316855, 1.25899173, -0.45533973]),
array([-0.57686454, -0.14026687, -0.10215045, -0.47547637, -0.72302332])]

```
==========聚类中心==========
[array([ 1.32175004, -0.57821878, -0.55743528, -0.96067712, -0.3459337 ]),
```

图表 24 求解：聚类中心（部分截图）

客户聚类结果：

R	F	M	C	L	prediction
7	140	293678	1.252314	2597	2
11	135	283712	1.254676	2615	2
97	23	281336	1.09087	2047	2
3	118	179514	1.398382	2714	2
65	22	234721	1.026085	2194	2
7	101	172231	1.386525	1355	2
2	64	169358	1.401596	2916	2
4	106	167113	1.369404	2070	2
74	20	222380	1.004904	1448	2

图表 25 求解：客户聚类结果（部分）

6.3.3 模型分类预测

这里使用数据清洗后的第一行数据进行客户类别分类，由结果可知该测试数据被模型分为第 2 类客户，与原聚类类别划分一致。

图表 26 类别分类测试

6.4 模型改进分析

6.4.1 模型优点

客户特征 LRFMC 模型方面，原 RFM 模型中 M 特征用一定时间内累计的飞行里程 M 与客户在一定时间内乘坐舱位对应的平均折扣率 C 来代替。同时考虑到了会员的入会时间在一定程度上能够影响客户的价值，增加客户关系长度 L 作为区分客户的一种特征，相对传统 RFM 模型覆盖客户特征更全，更适用于航空公司这一特定企业的客户特征分析。

聚类算法模型方面，K-means 算法简单易懂且聚类效果较好，通过手肘法合理的确定 K 值，增加模型的准确性。对于本文分析的航空公司客户数据，源数据 62988 条，清洗后的数据 3586 条，K-meas 聚类在数据量庞大的数据集相比传统系统聚类具有收敛速度快，聚类效率高的优点。

6.4.2 模型缺点

由于 K-Means 聚类算法是结果受初始值影响的局部最优的迭代算法，且受初始聚类中心点选择影响，若初始聚类中心为离群属性点，则会影响整体客户聚类分类效果。算法是不断迭代划分数据中心点的，噪声和离群点会干扰中心划分

的距离计算。

客户特征属性选择上暂未考虑航空客户因天气、行程等不确定因素导致的退票率，退票率也是航空公司研究客户质量的重要指标，因此模型还有待进一步改进。

6.4.3 模型改进

1. 聚类中心选择优化

目前已有 K-Means 聚类算法改进的相关研究，如 K-Means++算法在选取第 n+1 个聚类中心时，距离当前 n 个聚类中心越远的点会有更高的概率被选为第 n+1 个聚类中心，对 K-Means 随机初始化质心进行优化。二分 K-Means 算法首先将所有数据看做一个聚类，然后进行聚类划分，保证每一步得到的总体误差最小。

2. 距离计算优化

如 elkan K-Means 算法利用两边之和大于等于第三边,以及两边之差小于第三边的三角形性质，来减少距离的计算。对于一个样本点和两个质心，第一种思路是预先计算两个质心间的距离，若样本点到其中一个质心距离的两倍小于到另一个质心的距离，则该样本点直接属于该质心所属的簇；第二种思路是利用三角形的性质，得到样本到距离最短的质心。对 K-Means 算法的距离计算进行优化，加快迭代速度。

7 结果分析

每种客户五个特征的描述统计量如下图：

	R				F	...	C	L			
	count	unique	top	freq	count	...	freq	count	unique	top	freq
prediction						...					
0	944	377	402.0	13	944	...	673	944	742	699.0	9
1	890	279	4.0	30	890	...	22	890	669	2922.0	5
2	345	72	2.0	29	345	...	1	345	320	2479.0	3
3	701	383	4.0	12	701	...	109	701	575	454.0	5
4	706	242	4.0	18	706	...	14	706	548	820.0	7

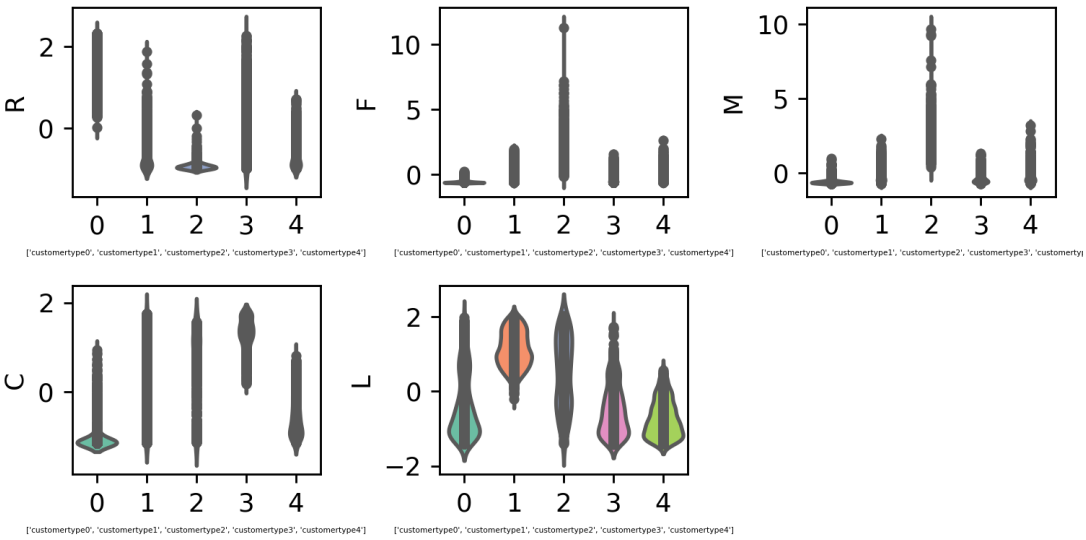
图表 27 聚类结果描述统计

客户群聚类 944 个样本，客户群 1 聚类 890 个样本，客户群 2 聚类 345 个样本，客户群 3 聚类 701 个样本，客户群 4 聚类 706 个样本，聚类数量分布较为均匀，可以看出聚类结果良好。

市场上常根据业务定义五个等级的客户类别：重要保持客户、重要发展客户、重要挽留客户、一般客户、低价值客户。下面将从不同维度分析聚类结果数据，并根据该标准对聚类的到的客户进行划分排名。

7.1 用户特征维度

本节对不同特征中与其他类别客户表现出明显差异的客户群进行分析。如图为得到的客户特征分布图，横坐标轴为客户聚类得到的不同客户类型：



图表 28 客户特征分布图（特征）

由图像可知 R 特征层面，客户群 0 的特征偏大，说明这一类客户上一次消费时间已经很久远，这些客户可能是已经或将要流失的客户。

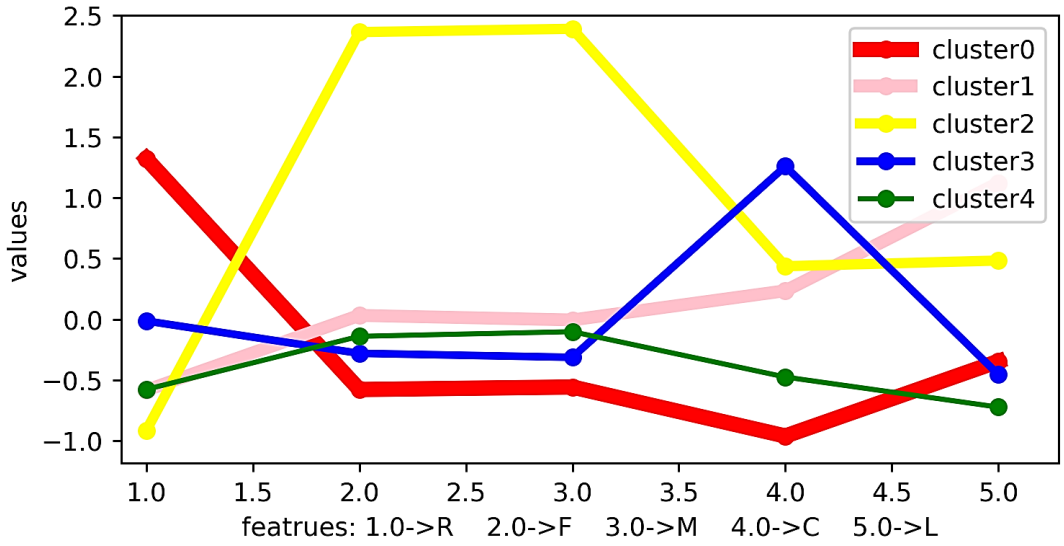
F 和 M 特征具有强相关性，因此一起分析，客户群 2 存在极高的异常点，分布极其偏态，客户群 2 可能是具有着消费频繁，消费金额大，即乘坐飞机次数多乘坐飞机的里程远的高价值客户群。

C 特征层面，客户群 3 具有较高的折扣系数，说明这类客户经常做折扣飞机，可以归类于经济性旅客或高消费高返利旅客。

L 特征层面，客户群 1 的特征偏大且较为集中，即入会时间早的老客户。

7.2 客户群维度

本节从聚类得到的不同客户类别维度分别对它们的特征进行分析。如图为得到的客户特征分布图，横坐标轴为客户聚类得到的不同客户类型：



图表 29 客户特征分布图（客户群）

由折线图可知客户群 0 在 R 处的值最大，且在 L,M,F,C 处的值都相对较小，可知客户群 0 距上次乘机时间比较久远，近期没有乘坐该航空公司的航班，可能是企业的重要挽留客户。

客户群 1 在 L 处特征最大，但在 R 处特征较小，其他特征适中，说明客户群 1 的入会时间较长，飞行频率高，是该航空公司的重要发展客户。

客户群 2 在 F,M 上的值最大，相对于其它类用户 C、L 特征也较大，且在特征 R 上的值最小，说明客户群 3 频繁乘机且近期都有乘机记录，对航空公司信任度和忠诚度都比较高，是有高价值的客户群。

客户群 3 在 C 处的值最大，在 F,M 处的特征值较小，说明客户群 3 是偏好乘坐经济舱位或高级舱位的客户群。

客户群 4 在所有特征上的值都很小，且在 L 处的值最小，说明客户群 4 是新入会员较多。

根据上述分析可以对不同客户的价值进行归类，定义 5 个等级的用户群体，根据其重要程度进行排列。

1. **重要保持客户**
    - 如上图中的客户群 2。
    - 这类客户最近一次乘机距今的时间长度（R）较低，飞行次数（F）或飞行里程（M）较高。
    - 是航空公司的高价值客户，是最理想的客户，即盈利价值最大的客户，航空公司应优先针对这类客户进行资源投放分配，进行差异化管理和一对一营销，提高这类客户的忠诚度和满意度。
2. **重要发展客户**
    - 如客户群 1。
    - 此类客户的入会时间（L）特征最大，但在最近乘机距今的时间长度（R）处特征较小，其他特征适中，说明客户群 1 的入会时间（L）较长，飞行频率高，是该航空公司的重要发展客户。
    - 是航空公司的潜在价值客户。航空公司要努力促使这类客户增加在本公司的乘机消费，增加积分和优惠券的发放。通过客户价值的提升，加强这类客户的满意度和粘性，提高他们转向竞争对手的转移成本，逐渐成为公司的忠诚客户。
3. **重要挽留客户**
    - 如客户群 0。
    - 此类客户在最近乘机距今的时间长度（R）处的值最大，且在飞行次数（F）或飞行里程（M）、入会时间（L）、平均折扣系数（C）处的值都相对较小，可知客户群 0 距上次乘机时间比较久远，近期没有乘坐该航空公司的航班，可能是企业的重要挽留客户。
    - 客户价值变化的不确定性很高。由于这些客户价值衰退的原因各不相同，所以掌握客户的最新信息，维护与客户的互动就显得尤为重要。航空公司应根据这些客户最近消费时间和消费次数的变幻推测客户消费的异动情况，并列出客户名单，对其重点联系，采用一定的营销手段，延长客户的生命周期。
4. **一般客户**
    - 如客户群 4。

- 这类客户的入会时间（L）相对最小，但是其他整体特征值较为平均，相对其他客户群时间长度（R）、飞行次数（F）、飞行里程（M）、平均折扣系数（C）都不突出，可知客户群 1 是正常乘坐航班的普通旅客群体且新用户较多。

5. **低价值客户**

- 如客户群 3。
- 这类客户的飞行次数（F）或飞行里程（M）较底，入会时间（L）、最近一次航班距今时间（R）较长，平均折扣率（C）最高，可能是偏好乘坐经济舱位或低频率出行选择高级舱位的客户群体。
- 是该航空公司的低价值用户，该航空公司对于这类用户可能可替代程度高，这类用户一般选择平均折扣率（C）较高的机票，且航班频率较低，对航空公司利益不大。

总结得到客户价值分析排名表：

表格 3 客户价值分类表

客户群	客户价值排名	解释
Cluster2	1	重要保持客户
Cluster1	2	重要发展客户
Cluster0	3	重要挽留客户
Cluster4	4	一般客户
Cluster3	5	低价值客户

7.3 营销建议

7.3.1 会员等级管理

根据本文客户聚类结果，航空公司的会员分为多个不同类型，可以将它们分别设置为不同等级会员。根据一定的标准计算得分可以将会员分为普通会员和核心会员。成为核心会员要求在一定时间内积累一定的飞行里程，达到得分要求后就能够在有效期内成为核心会员，并享受相应的高级别服务。有效期快结束时，根据相关评价方法确定客户是否有资格继续作为核心会员，然后对该客户进行相应的升级或降级。

另外还可以在此基础上进行促销，航空公司可以在对会员升级评价的时间点

之前，对那些接近但尚未到要求的比较高消费客户进行适当提醒升值采取一些加速升级或促销活动，刺激他们消费达到相应指标。这样既可以获得收益，同时也可以提高客户的满意度，增加企业核心会员规模。

7.3.2 积分兑换驱动

企业营销过程中常用首次兑换或者免费礼品吸引客户。在航空公司营销策略的制定中，可以令客户端里程或者航段积累到一定程度时才可以实现第一次兑换。据数据分析可知，该航空共公司的客户兑换积分的积极性不高。可以通过积分购票折扣吸引客户，刺激客户对该航空公司的用户忠诚度，并提高服务满意程度。另一方面可能存在部分客户不了解积分兑换模式的信息差问题，航空公司可以采取的措施时从数据库中提取出接近但是尚未达到首次兑换标准的会员，对他们进行短信、电话提醒或者促销。

7.3.3 捆绑联名销售

通过与银行发行联名卡、话费充值送积分、与其他电子商务平台合作发行票务套餐等措施来提高客户的粘性，时客户在其他企业的消费过程中获得本公司的积分，增强与公司的联系，扩大公司业务覆盖范围，增加航空公司在其他领域知名度，提高客户的忠诚度。增加用户粘性，维护客户关系，最大化生命周期内公司与客户的互动价值。

8 附录

8.1 参考文献

[1]. 胡海.基于 RFM 模型改进的企业营销平台客户价值分析[J].营销界,2022(15):11-13.

[2]. 王润清. 基于聚类分析的航空旅客在线购票行为研究[D].中国民航大学,2020.DOI:10.27627/d.cnki.gzmhy.2020.000311

[3]. 杨佳欣. 基于 RFM 改进模型 LFMN 的新浪微博用户价值分析[D].贵州财经大

学,2022.DOI:10.27731/d.cnki.ggzcj.2022.000329.
[4]. 闫春,刘璐.基于改进 SOM 神经网络模型与 RFM 模型的非寿险客户细分研究[J].数据分析与知识发现,2020,4(04):83-90.
[5]. 王长琼,邱杰,曹乜蜻,王艳丽.基于谱聚类算法的城市快递客户聚类研究[J].武汉理工大学学报：信息与管理工程版,2018,40(5):620-624
[6]. 李为康,杨小兵.一种改进的 RFM 模型在网店客户细分中的应用[J].中国计量大学学报,2020,31(01):85-91+134.
[7]. 韩世莲.基于客户动态需求属性的物流配送线路聚类优化[J].系统管理学报,2016,25(6):1146-1153
[8]. 朱沅海,林泉,万杰.一种结合 PSOA 的模糊 K-均值客户聚类算法[J].计算机工程与科学,2009,31(12):74-76
[9]. 叶苗群.基于混合 K—中心点的 Web 客户聚类[J].嘉兴学院学报,2005,17(3):54-56

8.2 参考网页资料

[1]. CSDN pyspark 基础
[2]. Python 数据分析案例 09——航空公司客户聚类分析
[3]. 航空公司客户价值分析——K-Means
[4]. 从 0 开始学 pyspark（十）：使用 pyspark.ml.clustering 模块对商场顾客聚类
[5]. 基于 RFM 的航空公司客户价值聚类分析
[6]. k-means 聚类的原理以及缺点及对应的改进

8.3 文件说明

8.3.1 Csv 文件

[1] type_result.csv：客户聚类结果及对应特征文件；
[2] data_LRFMC_standard.csv：客户特征标准化 dataframe；
[3] data_LRFMC.csv：客户特征 dataframe；
[4] type_des.csv：分类结果数据描述；

[5] des_air_data.csv：源数据描述。

8.3.2 Python 源文件

[1]. Pre_analyse.py:源数据文件探索性分析、源数据分析可视化；

[2]. K_means.py:kmeans 聚类求解，聚类结果可视化；

[3]. Pre_resolve:数据清洗，数据预处理，聚类数探索。