基于K-means的电商客户聚类研究

[摘要]电子商务平台的迅猛发展推动了经济的增长，同时也为消费者提供了极大的便捷，但是伴随着信息大爆炸，电子商务的平台数据呈现指数型增长，积累了成千上万的用户信息和海量的购买行为的数据，如何从这些数据挖掘有价值的数据，从中识别不同的客户群体以此制定针对性的营销策略，成为许多电子商务平台面临的难题，因此本文基于UCI的平台的电商零售数据集对客户群体进行细分。

本文对于电商零售数据集进行数据集预处理后，基于RFM模型理论分别计算出R值、F值和M值，根据三种常见的聚类指标：Davies-Bouldin指数、Calinski-Harabasz指数和轮廓系数求出合适的聚类数目后运用K-means算法将客户群体进行细分，研究发现该企业的客户群体可以划分为四类，其中一般挽留客户和重要保持客户占比高。因此，该企业应针对此两类采取差异化策略，以更有效地提升客户活跃度和忠诚度。

目录

[第一章 绪论 4](#_Toc199404566)

[1.1 研究背景和研究意义 4](#_Toc199404567)

[1.1.1 研究背景 4](#_Toc199404568)

[1.1.2研究意义 4](#_Toc199404569)

[1.2 研究内容 5](#_Toc199404570)

[第二章 数据预处理与数据探索 6](#_Toc199404571)

[2.1 数据集说明 6](#_Toc199404572)

[2.1.1 电商零售数据集 6](#_Toc199404573)

[2.2 电商零售数据预处理 7](#_Toc199404574)

[2.2.1 缺失值处理 7](#_Toc199404575)

[2.2.2 无效数据处理和重复值处理 8](#_Toc199404576)

[2.2.3 相关指标的计算 10](#_Toc199404577)

[2.3 电商零售数据探索 12](#_Toc199404578)

[2.3.1 整体最近购买时间的情况分析 12](#_Toc199404579)

[2.3.2 整体购买次数的情况分析 14](#_Toc199404580)

[2.3.3 整体消费金额的情况分析 15](#_Toc199404581)

[第三章 基于K-MEANS的客户细分模型 17](#_Toc199404582)

[3.1 数据标准化处理 17](#_Toc199404583)

[3.2 K-MEANS模型构建 17](#_Toc199404584)

[3.3 客户等级划分 20](#_Toc199404585)

[第四章 措施和建议 22](#_Toc199404586)

# 第一章 绪论

## 1.1 研究背景和研究意义

### 1.1.1 研究背景

本文借助机器学习的相关技术方法，以 UCI 平台的两份数据集作为研究案例，以客户细分作为理论基础，对电商客户群体进行精细化划分，识别特征各异的消费群体，并最后针对不同客户群体提出差异化的营销策略建议，帮助企业更精准地匹配用户需求。

### 1.1.2研究意义

对于电商客户细分，具有以下研究意义：

1、精准营销和个性化推荐，客户细分有助于识别具有不同需求和行为特征的客户群体，通过有效的客户细分，平台能够更准确地匹配用户需求，推送更相关的产品信息，进而提高用户的购买意愿和满意度。

2、优化资源分配，通过了解哪些群体贡献了主要收入，平台能够在预算分配和客户管理上更具针对性，从而实现资源的最优化配置。

## 1.2 研究内容

本文的研究方法主要基于数据挖掘和机器学习技术，具体方法分为以下几个部分：

1、数据预处理：在正式建模前，使用Python对 UCI 平台的电商零售数据集进行数据预处理，包括缺失值填补、无效数据清洗、重复值处理等。

2、数据探索分析：在数据预处理后，在SPSS中采用数据可视化和统计分析方法对数据进行探索性分析，包括对分类变量和连续性变量的描述。

3、客户细分模型构建：基于电商零售数据，采用 K-Means 聚类算法对客户群体进行细分。通过数据标准化处理后构建聚类模型，根据购买频率、消费金额、最近购买时间指标，将客户划分为不同层级。

# 第二章 数据预处理与数据探索

## 2.1 数据集说明

### 2.1.1 电商零售数据集

该数据集是来自UCI Machine Learning平台(<https://archive.ics.uci.edu/dataset/502/online+retail+ii>)，记录了英国注册的在线电商零售公司在2009年12月1日至2011年12月9日的所有客户交易情况，包含100多万条交易记录，一共有8个变量，如下表所示。

|  |  |  |
| --- | --- | --- |
| 变量 | 描述 | 数据类型 |
| InvoiceNo | **发票编号，每个编号包含6位数字编号，以”C”开头为取消交易** | **标称型** |
| StockCode | **商品编号，包含5位数字编号** | **标称型** |
| Description | **商品描述名称** | **标称型** |
| Quantity | **商品数量** | **数值型** |
| InvoiceDate | **发票的日期和时间** | **数值型** |
| UnitPrice | **商品的单价** | **数值型** |
| CustomerID | **客户编号，包含5位数字编号** | **标称型** |
| Country | **客户所在国家名称** | **标称型** |

表 2.1.1 电商零售数据集描述

## 2.2 电商零售数据预处理

### 2.2.1 缺失值处理

利用Python的pandas库读取文件并查看该数据集结构，返回数据列、数据类型、非空值计数等关键信息。

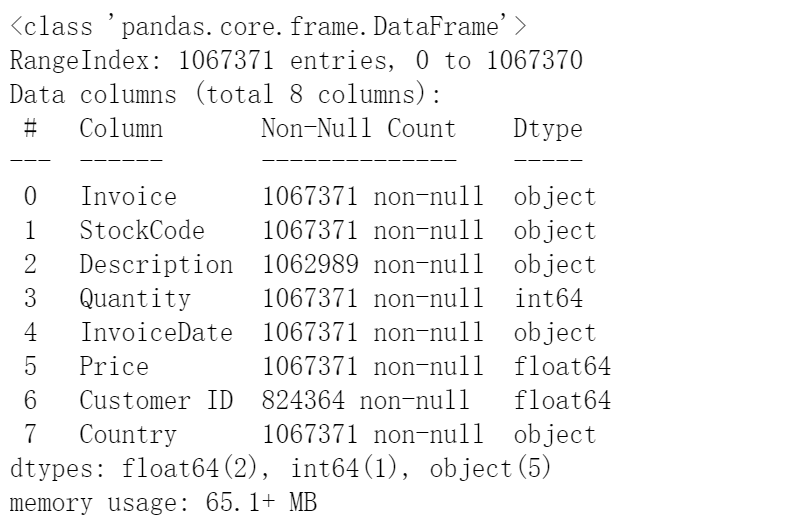
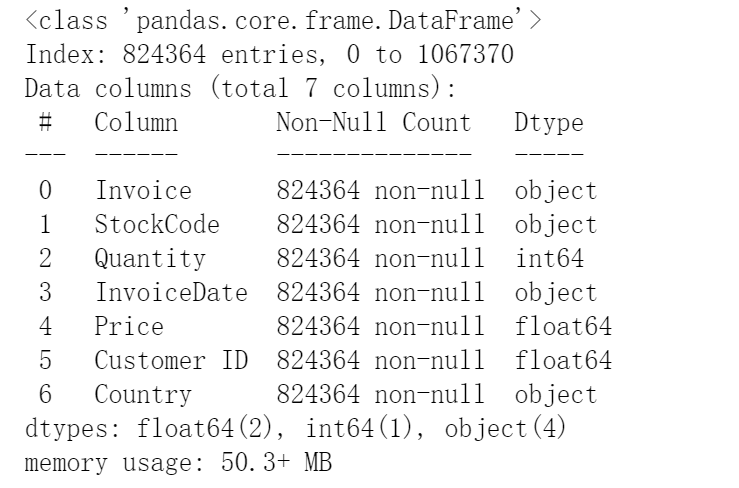


图 数据集基本结构

从上图可以看出，该数据集共包含 1067371 条记录。大多数列没有缺失值，但 Customer ID 列缺失较为严重，约有 20 多万条记录的编号存在缺失值，同时Description 列有少量缺失值，考虑到后续的聚类分析需要为每个用户计算相关指标，因此客户编号是必须存在的，对于 Customer ID 存在null值的记录需要进行剔除。

由于 Description 列主要用于描述商品，该变量与后续分析无关，可以选择保留或剔除该列，本文剔除该列。

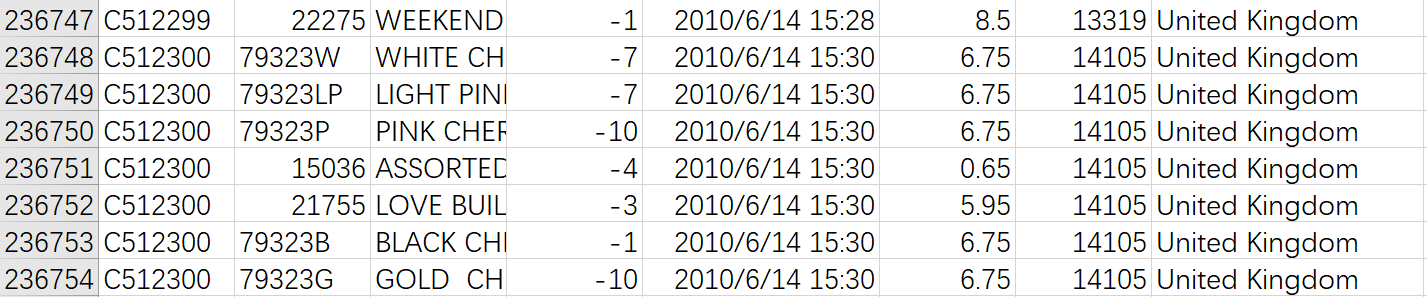
处理后，该数据集一共有824364条记录，包含7个变量。



### 2.2.2 无效数据处理和重复值处理

根据数据集的变量说明，InvoiceNo 列中以 “C” 开头的值表示取消的订单，经统计一共有18744 条取消订单记录，这些订单对于构建 RFM 模型来说是没有任何意义的，因此为了确保数据的有效性，需要剔除此类订单。

同时，StockCode列是包含5位数字的编号，但某些数据包含字母，属于无效数据，因此需要剔除包含字母的记录。



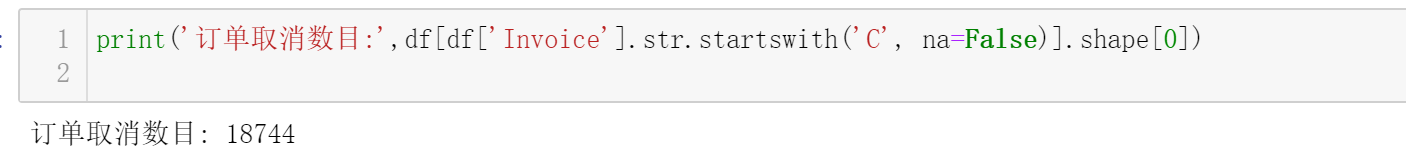


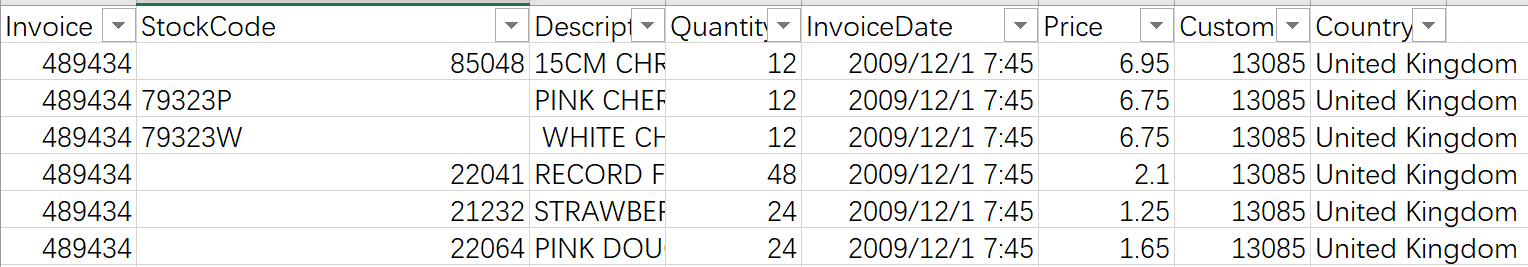
图 以C为开头的编号的发票订单



图 StockCOde列包含无效数据值

进一步观察数值型发现，部分数据中Price列的值为0，可能是该商品是赠品或奖品，为保证能正确统计客户真实消费行为，因此对于Price为0的记录需要进行剔除，剔除后有效记录是713989条。

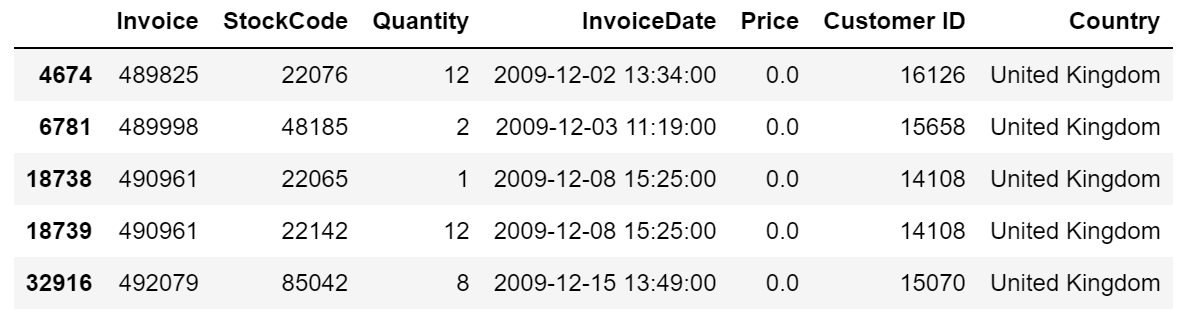


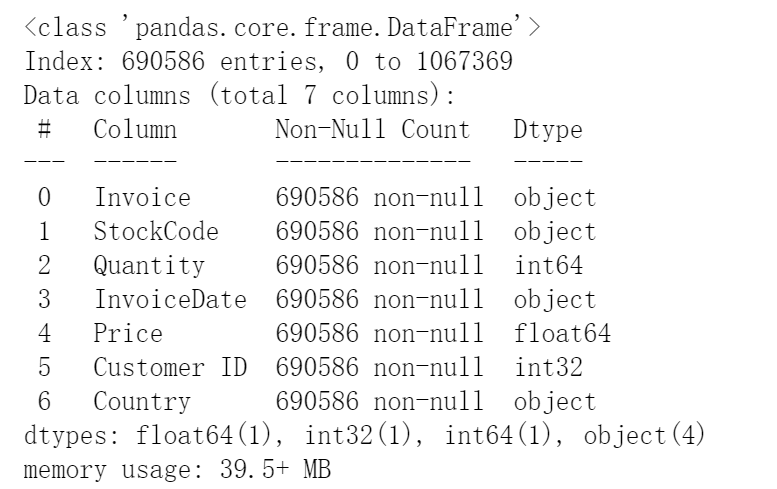
图 数值型的数据统计和Price为0的记录

同时该数据集中存在20403条完全相等的重复记录，为了避免重复汇总计算，应给予剔除，仅保留第一次出现的记录。



图 重复行数统计

上述数据预处理后，总共的有效记录是690586条。



### 2.2.3 相关指标的计算

由于本文主要是研究RFM模型，但该数据集并没有包括R(最近一次购买的时间)，F(购买次数)，M(总共消费次数)，因此需要从原始数据中计算相关指标。

首先，对于 R 值的计算，需获取该数据集中最晚的交易日期，因此确定日期界限，经过统计，数据集中最晚的交易日期为 2011-12-09 12:50，日期界限需要大于等于该时间点，本文选择 2011-12-10 00:00 作为日期界限。 

图 日期的最大值

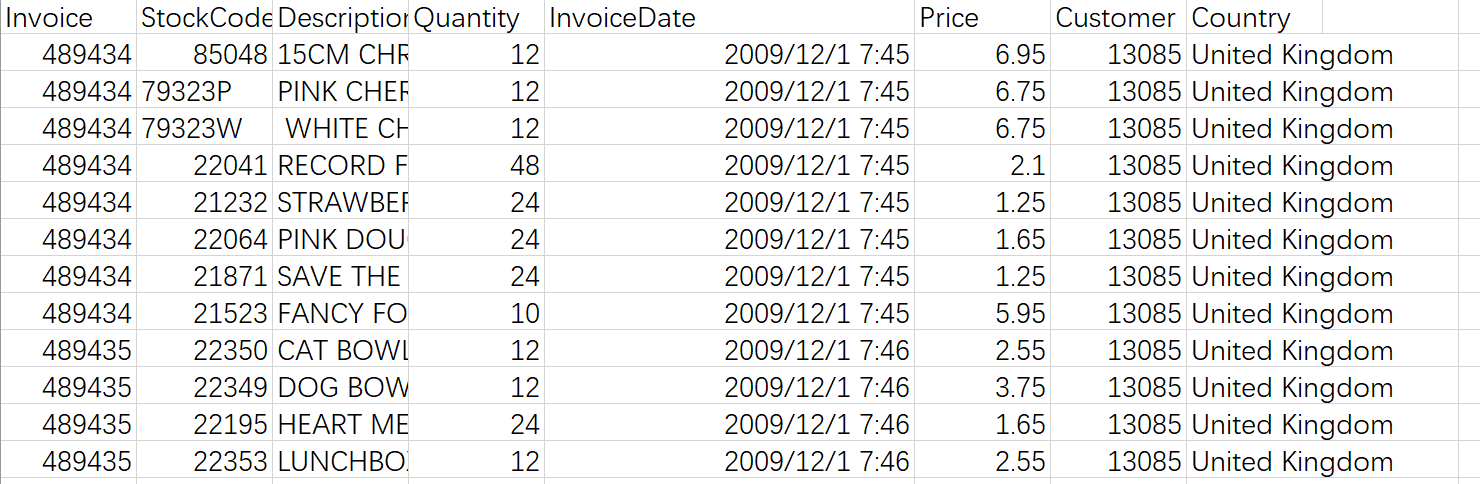
由于同一客户在不同时间段可能有多次消费记录，因此需要选取客户最近一次的购买记录。以客户编号 13085 为例，该客户存在订单编号 489434 和 489435 这两条记录，应选取最近的一次消费时间，即 2009-12-1 7:46，作为该客户的最近购买时间。将界定的日期界限与客户购买时间进行相减按照天数为单位即为R值。

图13085编号的客户

关于F值的主要是统计每个客户总的消费次数，每个客户每次进行一笔交易，会得到唯一的发票编码用于识别该笔交易，因此只需要统计每个客户不同发票编码数量即为消费次数。在Python中使用pandas库的groupby实现统计。如下图。



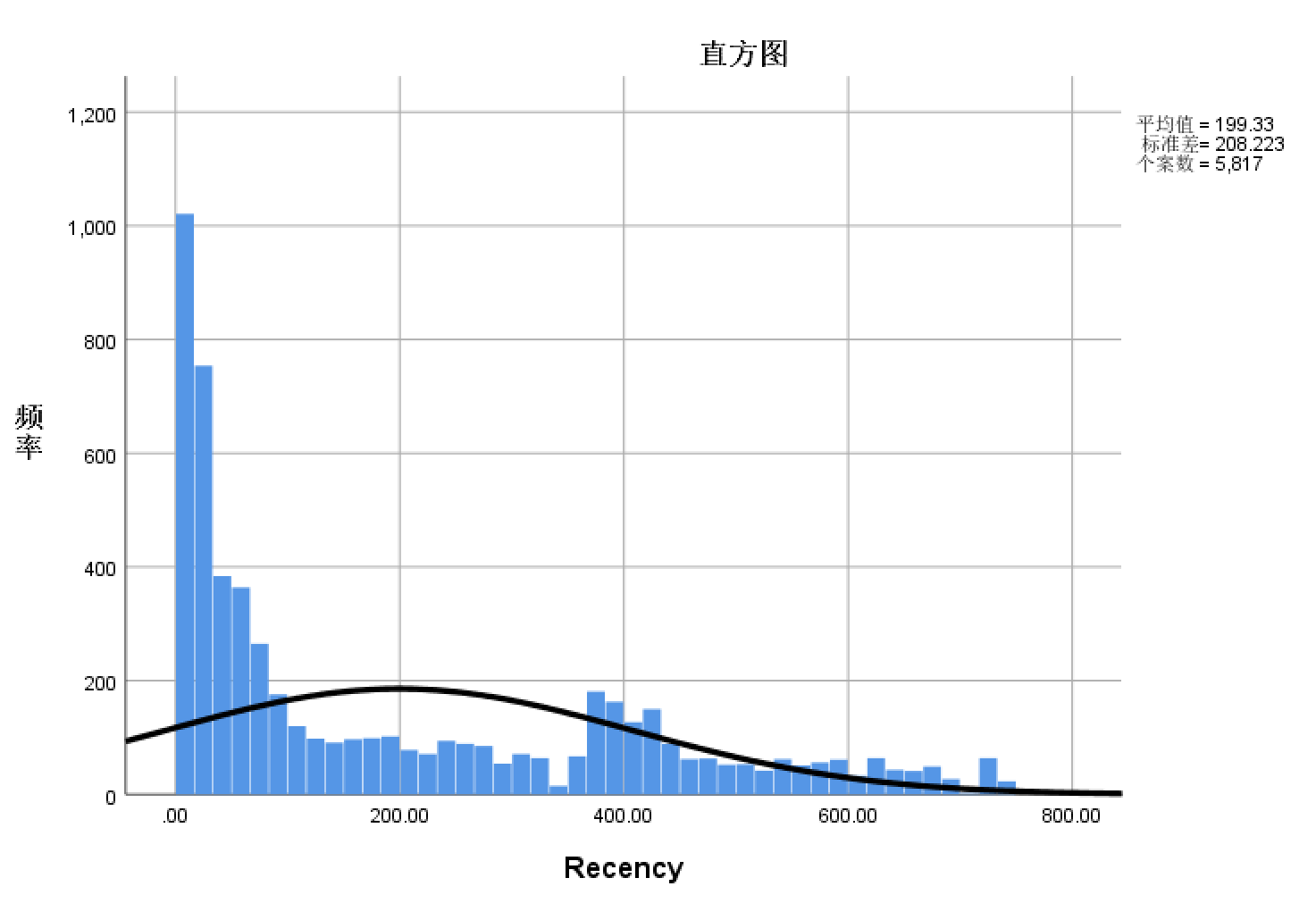
图 groupby函数实现F值统计

关于M值的计算涉及到每个客户的总消费金额，在原始变量中存在商品的Quantity和Price列，因此每一笔交易金额即为Quantity\*Price,对每个客户的交易记录的交易金额进行相加可得到每个客户最终的总消费金额。最终通过上述计算得到5817名客户的相关指标。

## 2.3 电商零售数据探索

### 2.3.1 整体最近购买时间的情况分析

对数据预处理后的5817名客户在SPSS中进行分析，得到如下用户整体最近购买时间的直方图。



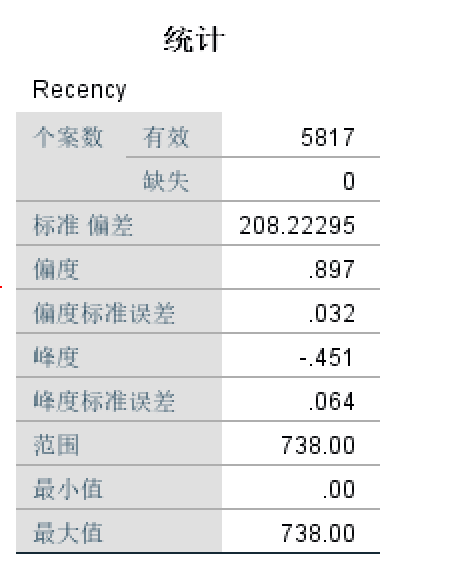
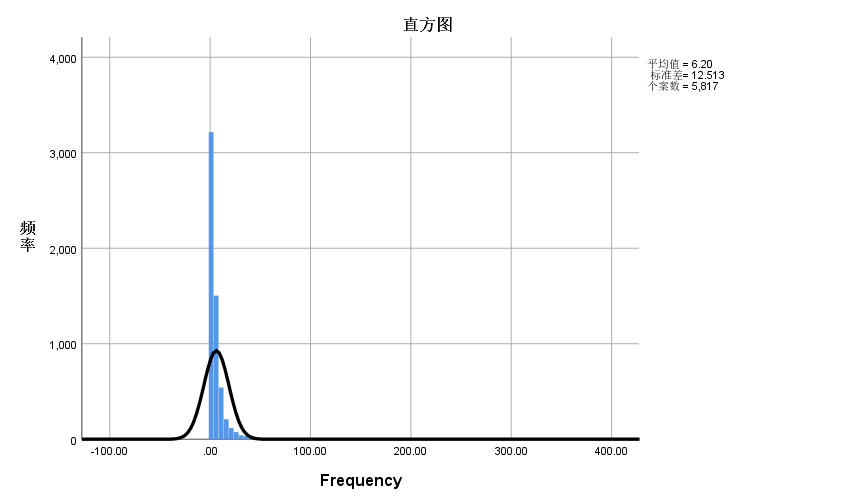


图 整体最近购买时间

可以看到直方图分布呈现右偏情况，偏度系数为0.897，属于中等偏态，其平均值为199.33，表明大多数客户的最近一次购买时间集中在平均值以下，客户活跃度较高。但少数客户的购买时间离当前时间较远，其峰度为-0.451，标准差相对较大，说明数据分布相对较平坦，没有明显的尖峰，大多数客户的购买行为较为分散，且没有集中在某一个固定的时间段。

### 2.3.2 整体购买次数的情况分析



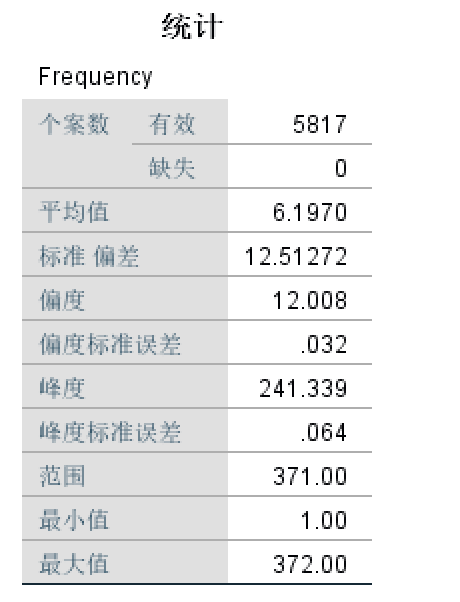


图 整体购买次数

从上图可以看到购买频数分布呈现严重的右偏情况，偏度为12.008，虽然其平均值为6次购买，但大多数客户的购买次数要低于这个值，且从峰度看购买频率分布呈现极高的尖峰，为241.339，说明绝大多数客户的购买次数集中在一个较低的频次，但有极少数客户购买次数异常多。有极少数客户购买次数非常多，最大为372 次，远远超出平均水平。说明这些客户是企业的忠实核心用户，对企业的整体销售贡献较大。

### 2.3.3 整体消费金额的情况分析

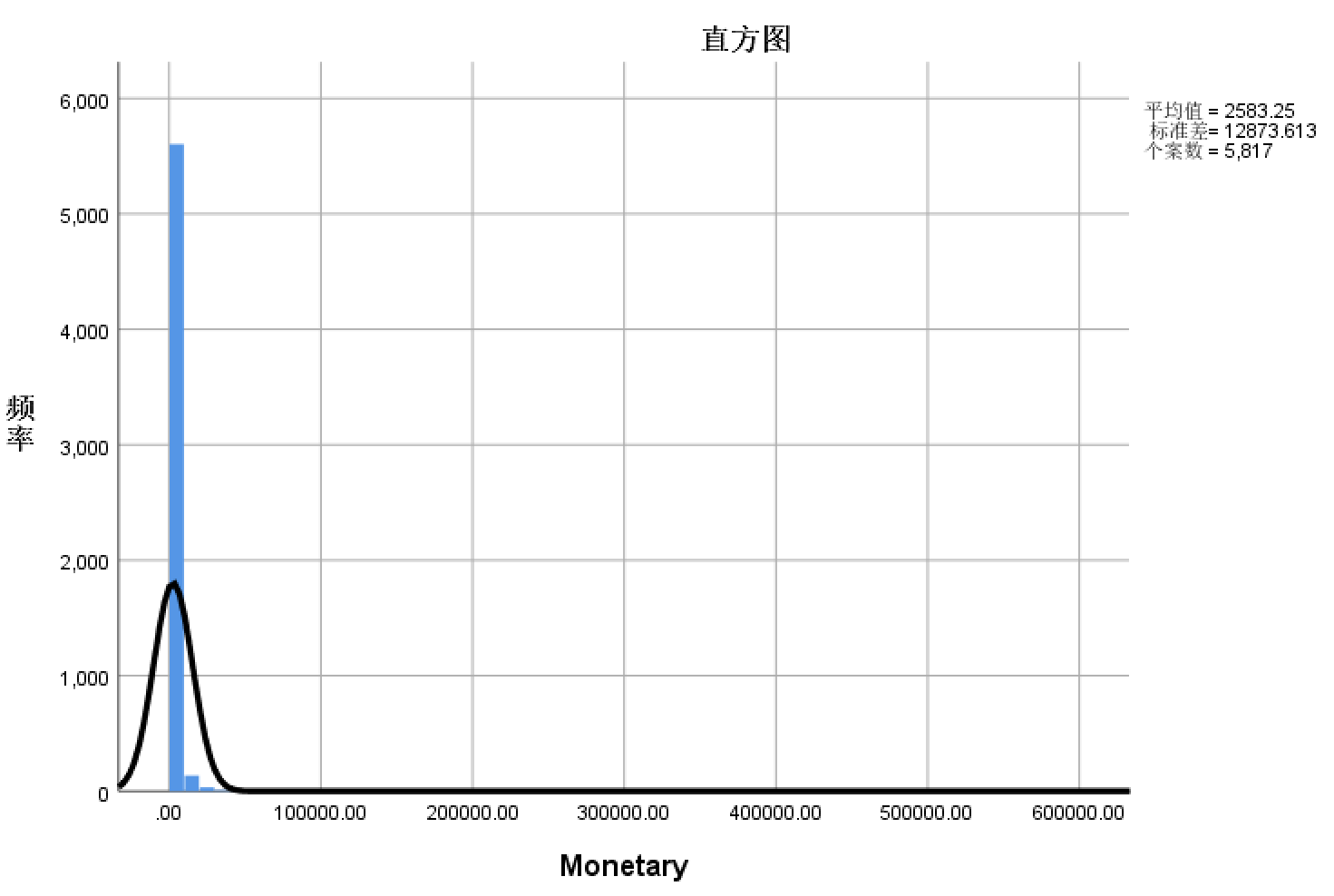




图 整体消费金额情况

从图中可以看到相较于前两个指标，客户的消费金额差异显著，偏度为 26.025，表明消费金额分布严重右偏，大部分户的消费金额较低，低于平均值2583.25，峰度为 885.527，说明数据分布的集中度极高，大多数客户的消费金额集中在低值区间，少数客户的消费金额非常高，最高达到了525573，形成了尖峰现象。

​

# 第三章 基于K-MEANS的客户细分模型

## 3.1 数据标准化处理

观察电商零售数据发现，该数据集中的Recency、Frequency和Monetary列的数值存在较大的差异，因此需要对这三列进行标准化处理。

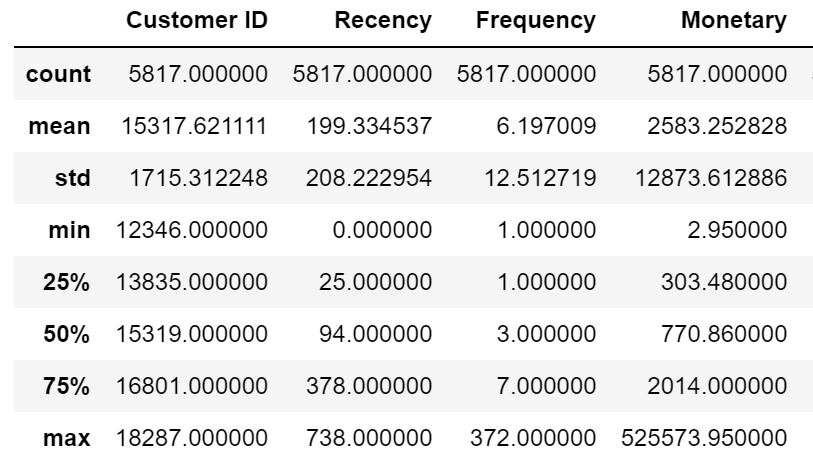


图 数据集结构

所使用的公式如下所示，其中X是原始数据，是数据的均值，是数据的标准差，该公式可以将数据的每个特征缩放，使其均值为0，标准差为1，有助于消除特征之间的量纲差异。

图 标准化公式

## 3.2 K-MEANS模型构建

对数据集进行标准化处理后，K-MEANS算法需要确定适合的聚类数。因此本文选取了三种常用的聚类效果评估指标：Davies-Bouldin指数(DBI)、Calinski-Harabasz指数(CHI)和轮廓系数（Silhouette Score）对不同的聚类数进行评估。由于CHI值较大，本文将其缩小5000倍进行绘画，得到如下图。

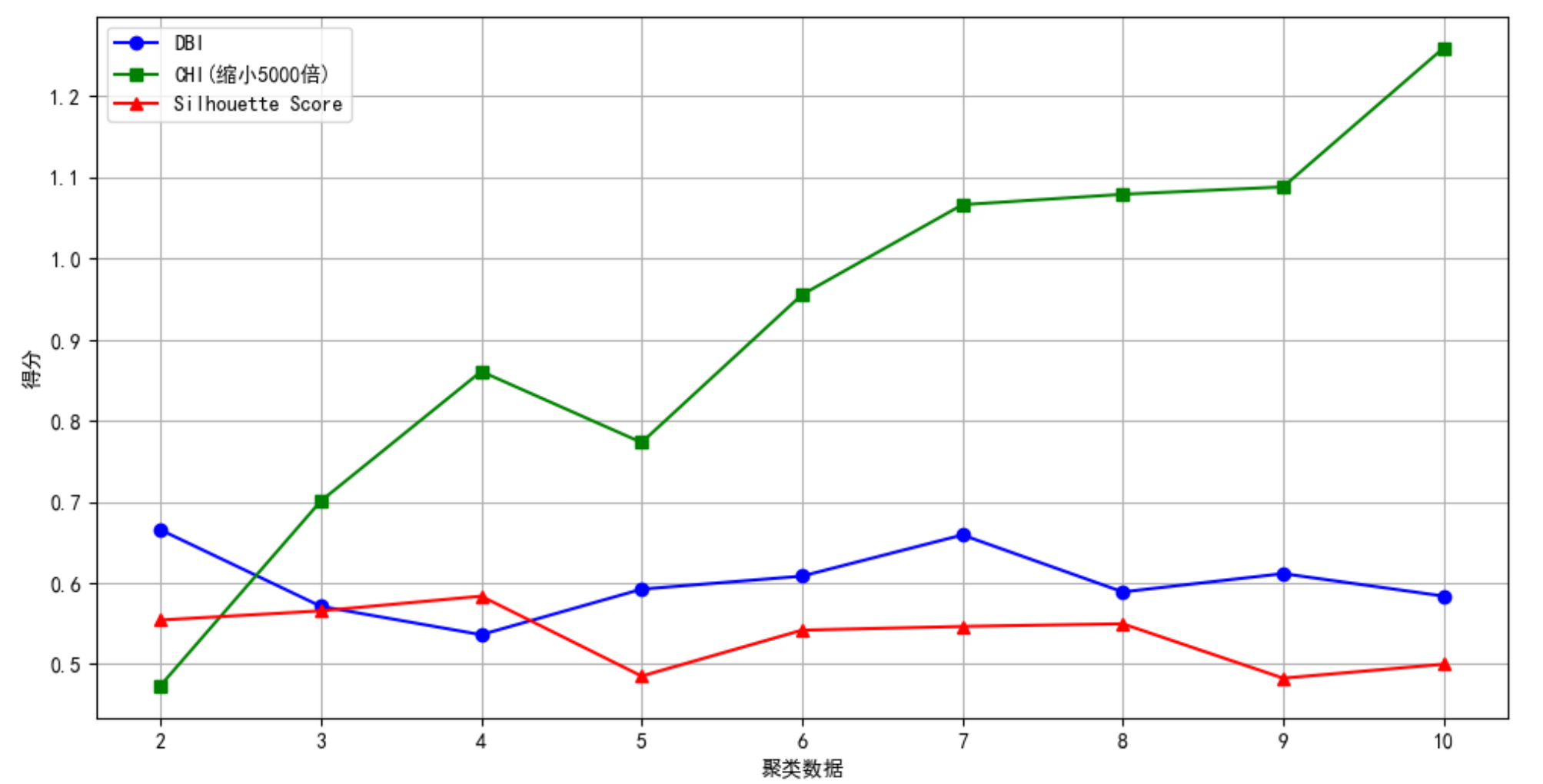


图5.2.1 三种聚类评估指标的表现

观察上述图可以看到随着聚类数目的增加，CHI指数不断上升，表明簇内紧密度较高，簇间分离度较大。然而在聚类数为4和5时，CHI指数出现了下滑的趋势，虽然在聚类数超过5后CHI指数不断上升，得分越来越高，但后面的聚类数目很有可能会导致出现模型拟合过度的情况。结合DBI和轮廓系数的变化，可以发现DBI在聚类数为4时达到最小值，为0.5363，而此时的轮廓系数接近0.6，表明此时的聚类效果较好，簇内紧密且簇间分离明显。基于以上三种指标的综合考虑，最终确定聚类数为4类。

本文使用Python的sklearn库中的K-means算法对数据集进行聚类，聚类的结果如下图所示。

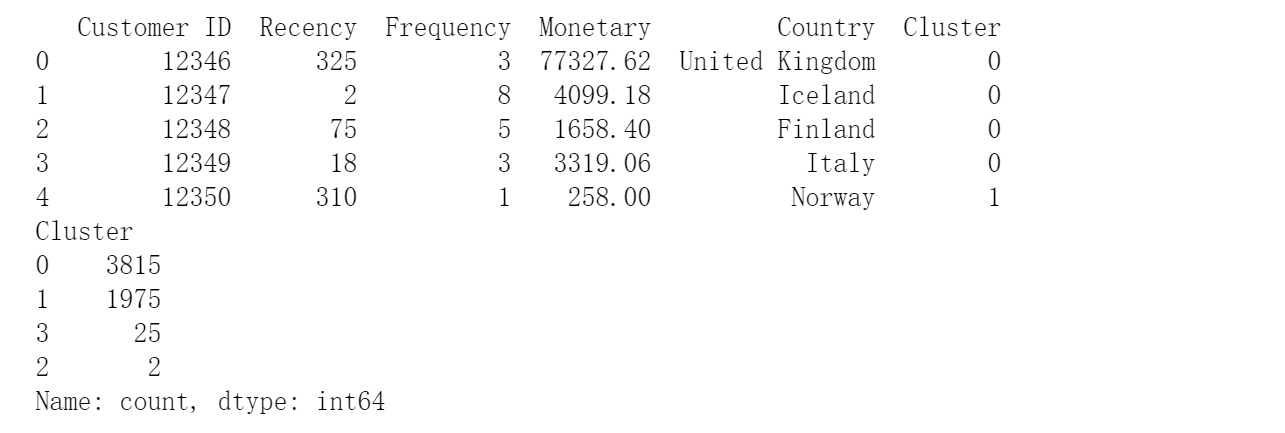


图 聚类结果展示

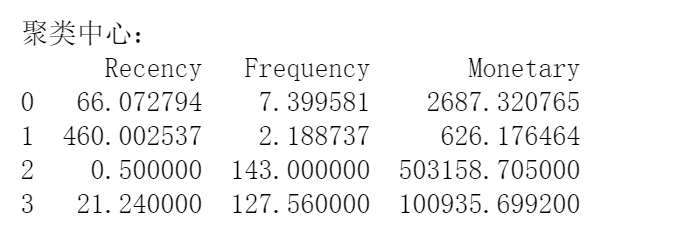
从图中可以看出，簇0 和簇1 的客户数量分别为3815和1975，表明该电商零售公司的主要客户群体集中在这两类客户中，占据了绝大部分比例，但是簇2和簇3的客户数量相对较少，总共仅有27人，远远不足100人，说明这些客户与其他簇的客户存在显著差异，簇2和3是特殊类型的客户群体，需要引起企业的重视。

进一步观察聚类的质心发现，每个簇之间的特征存在较大的差异性，簇0的客户最近一次购买时间为66天前，距离日期界限不算太远，说明客户活跃度适中，每个客户的购买频率为7次购买，消费频次也处于中等水平。平均消费金额为2687.32元，表明这些客户的消费水平一般。

簇1相比于簇0来说三个指标都偏低，尤其是最近一次的购买时间距离长达460天，总消费金额为626，表明这些客户已经很久没有购买，属于低活跃度、购买频率较低、消费金额也较低的客户群体。

同时可以发现簇2的客户最近一次购买时间距离几乎为0，说明这些客户刚刚进行了购买，客户的购买频率非常高，平均购买次数高达143次，表明这些客户是超高活跃的客户，平均消费金额极高，达到503,158.71元，远高于其他簇。

相比簇2，簇3的各项指标虽然有所差距，但仍属于高活跃度、高购买频次、高消费金额的客户群体。客户平均消费次数为127.56次，消费金额为100,935.70元，说明这类客户也是企业的重要客户群体。



## 3.3 客户等级划分

根据RFM理论模型，RFM模型将客户群体按照“低“，”高“两类情况对客户分为8大类。可以将上述4类结果进行客户群体，分类结果如下所示。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 簇类 | R值 | F值 | M值 | 客户等级 |
| 0 | 低 | 高 | 高 | 重要保持客户 |
| 1 | 低 | 低 | 低 | 一般挽留客户 |
| 2 | 极高 | 极高 | 极高 | 重要价值客户 |
| 3 | 高 | 极高 | 极高 |

表 客户群体划分

由上图客户等级可以看到，该企业的大部分客户群体集中在重要保持客户和一般挽留客户中。其中重要保持客户人数最大，为3815人，簇2和簇3的客户群体是企业的高价值核心客户，这类客户对企业的贡献最大。

# 第四章 措施和建议

针对上述第五章的电商平台客户群体的细分，企业应根据不同客户的特征采取差异化的策略，以最大化客户价值，企业可以采取如下建议：

1、对于重要保持客户，这些客户有较高的消费频次和金额，但活跃度较低，因此企业应通过定期推送个性化推荐、专属折扣以及实施会员忠诚计划来维系客户关系，防止流失。

2、对于一般挽留客户，由于这些客户的消费频次和金额均较低，且活跃度不高，企业可以通过激励性的促销活动和定期购物提醒来重新激活他们。

3、对于重要价值客户，这些客户频繁购买且消费金额极高，是企业的核心客户群体，可以提供个性化服务、专属客服和快速退换货等优待，增强其忠诚度。