****

**Python数据处理**

**实验报告**

|  |  |
| --- | --- |
| 学生姓名 | 李栩雅 |
| 学 号 | 8208201116 |
| 专业班级 | 计算机科学与技术2002 |
| 指导教师 | 王 斌 |
| 学 院 | 计算机学院 |
| 完成时间 | 2022.5.30 |

计算机学院

2022年5月

目录

[实验2：航空公司客户价值分析 3](#_Toc104969253)

[一、 实验目标与基本要求 3](#_Toc104969254)

[1. 实验目标 3](#_Toc104969255)

[2. 基本要求 3](#_Toc104969256)

[二、 实验过程设计 3](#_Toc104969257)

[1. 实验理论 3](#_Toc104969258)

[2. 实验教学过程 4](#_Toc104969259)

[三、 程序概要设计 4](#_Toc104969260)

[1.思维导图 4](#_Toc104969261)

[2.主要算法介绍 5](#_Toc104969262)

[具体步骤 5](#_Toc104969263)

[四、 主要程序分析 6](#_Toc104969264)

[五、 运行结果展示 9](#_Toc104969265)

[六、 实验结果分析 11](#_Toc104969266)

[七、 模型应用 12](#_Toc104969267)

[八、 心得体会 15](#_Toc104969268)

实验2：航空公司客户价值分析

# 实验目标与基本要求

### 实验目标

重点结合航空公司客户价值分析的案例介绍K-Means聚类算法在客户价值分析中的应用。此外，介绍基于RFM客户价值分析模型的不足，使用K-Means算法构建航空客户价值分析LRFMC模型，详细的描述数据分析的整个过程。

### 基本要求

1. 了解RFM模型的基本原理。
2. 掌握K-Means算法的基本原理与使用方法。
3. 比较不同类别客户的客户价值，制定相应的营销策略。

# 实验过程设计

### 实验理论

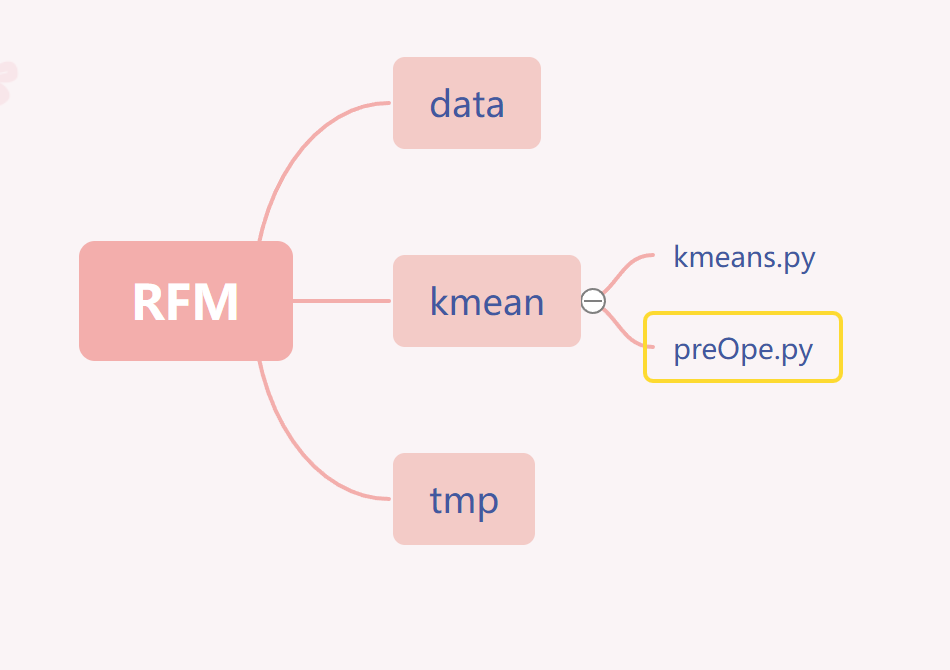
1. 分析航空公司现状。
2. 认识客户价值分析。
3. 熟悉航空客户价值分析的步骤与流程。
4. 处理缺失值与异常值。
5. 构建爱你航空客户价值分析关键特征。
6. 标准化LRFMC 5个特征。
7. 了解K-Means聚类算法。
8. 分析聚类结果。
9. 模型应用。

### 实验教学过程

1. 处理数据缺失值与异常值。
2. 构建航空客户价值分析的关键特征。
3. 标准化LRFMC 5个特征。
4. 构建K-Means聚类模型。
5. 评价K-Means聚类模型。

# 程序概要设计

## 1.思维导图



1. data文件夹存储本次实验的数据，使用.csv保存
2. kmean保存本次实验的程序，其中preOpe是进行一些数据的前置处理，kmeans则是对数据进行聚合并可视化
3. tmp是一个以numpy形式压缩的数据包，在程序中会使用

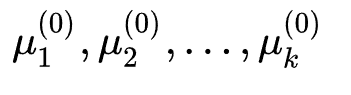
## 2.主要算法介绍

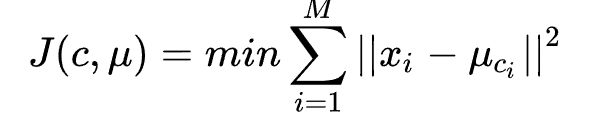
**k均值聚类算法（k-means clustering algorithm）是一种**[**迭代**](https://baike.baidu.com/item/%E8%BF%AD%E4%BB%A3/8415523)**求解的聚类分析算法，其步骤是，预将数据分为K组，则随机选取K个对象作为初始的聚类中心，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个**[**聚类**](https://baike.baidu.com/item/%E8%81%9A%E7%B1%BB/593695)**。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，**[**误差**](https://baike.baidu.com/item/%E8%AF%AF%E5%B7%AE/738024)[**平方和**](https://baike.baidu.com/item/%E5%B9%B3%E6%96%B9%E5%92%8C/783894)**局部最小。**

## 具体步骤

KMeans的核心目标是将给定的数据集划分成K个簇（K是超参），并给出每个样本数据对应的中心点。具体步骤非常简单，可以分为4步：

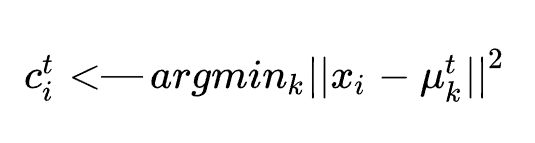
（1）数据预处理。主要是标准化、异常点过滤。

（2）随机选取K个中心，记为 

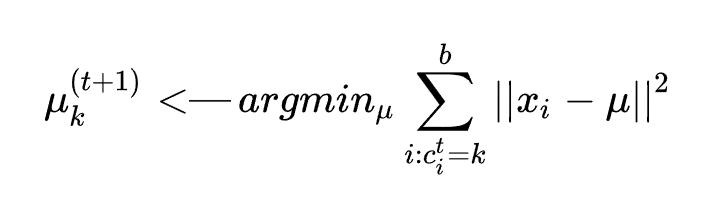
（3）定义损失函数： 

（4）令t=0,1,2,... 为迭代步数，重复如下过程知道  收敛：

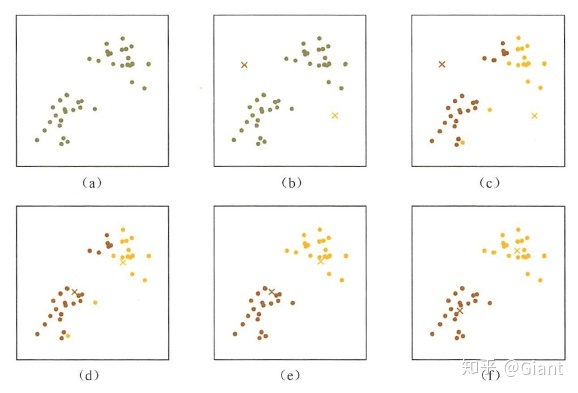
（4.1）对于每一个样本 ，将其分配到距离最近的中心



（4.2）对于每一个类中心k，重新计算该类的中心



**KMeans最核心的部分就是先固定中心点，调整每个样本所属的类别来减少  ；再固定每个样本的类别，调整中心点继续减小 。两个过程交替循环，  单调递减直到最（极）小值，中心点和样本划分的类别同时收敛。**

KMeans迭代示意图

# 主要程序分析

**preOpe.py**

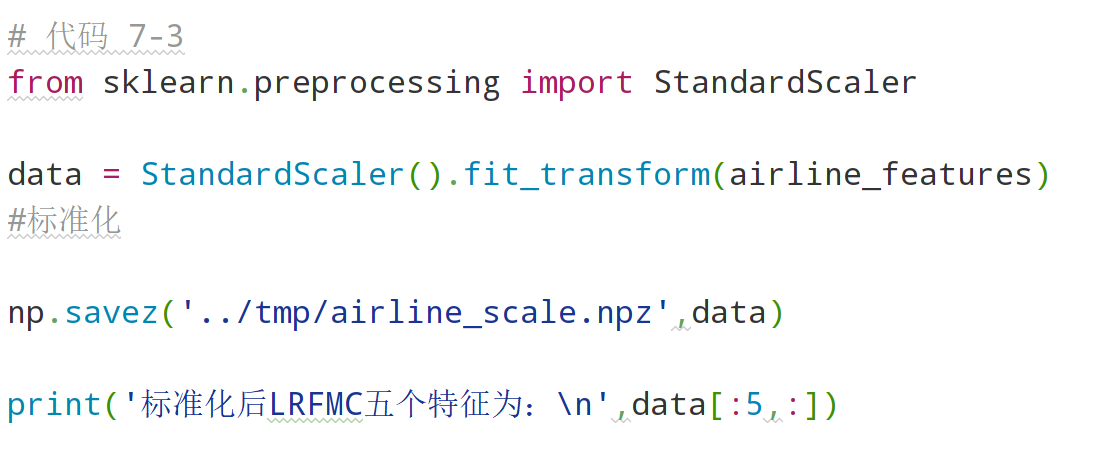


这里是去掉表格中的异常数据，便于后续处理分析



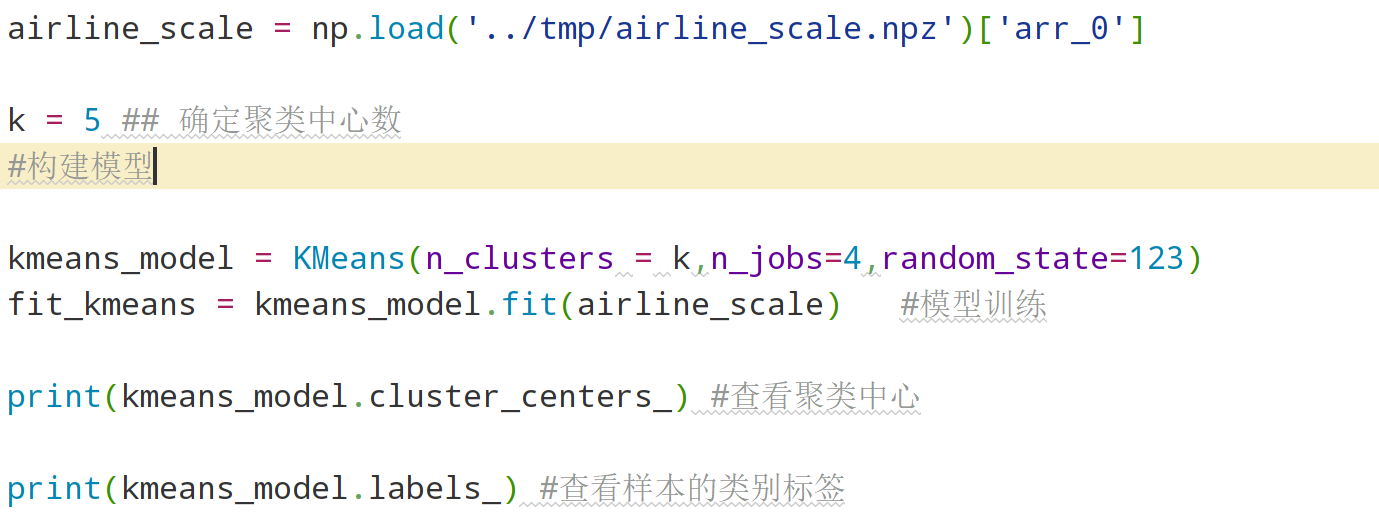
此处将客户关系长度L、消费时间间隔R、消费频率F、飞行里程M和折扣系数的平均值C这5个特征作为航空公司识别客户价值的特征，记为LRFMC模型。

原始数据中属性太多，根据航空公司客户价值LRFMC模型，选择与LRFMC指标相关的6个属性：FFP\_DATE、LOAD\_TIME、FLIGHT\_COUNT、AVG\_DISCOUNT、SEG\_KM\_SUM、LAST\_TO\_END。删除与其不相关、弱相关或冗余的属性。



这里使用StandardScaler通过删除平均值并缩放到单位方差来对数据进行标准化。

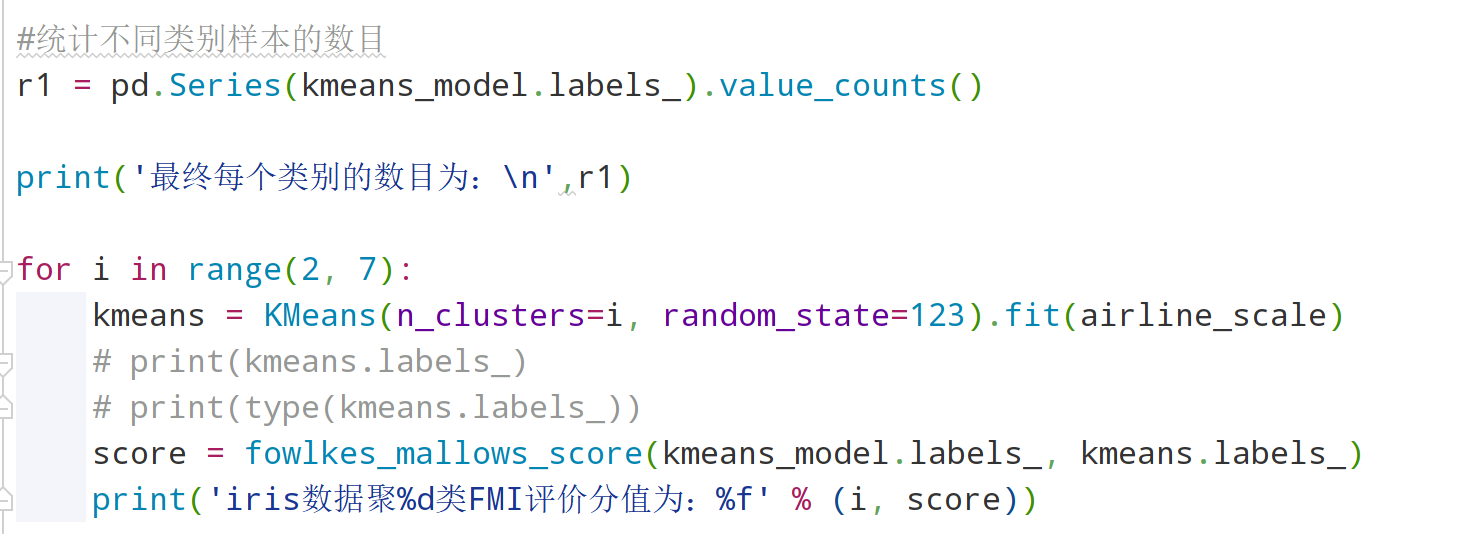
**kmeans.py**



使用Kmeans算法将数据分类聚合，使用airline\_scale进行模型训练



绘制可视化雷达图

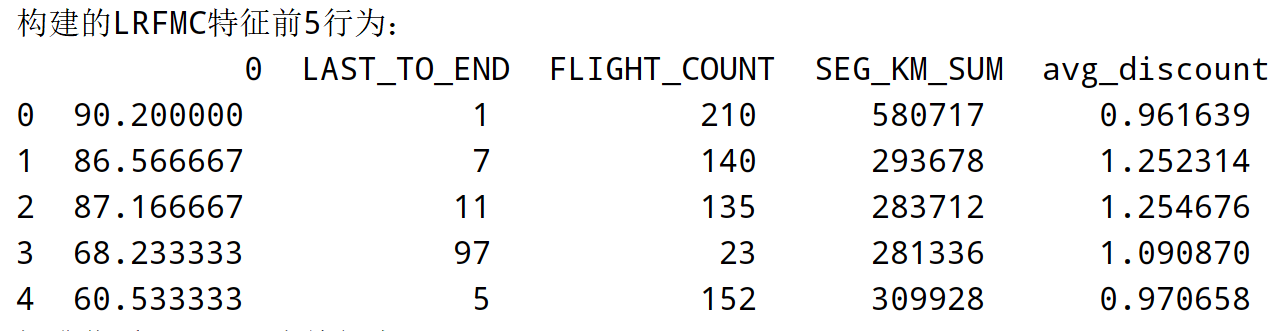


统计不同样本的数目并评估聚类模型

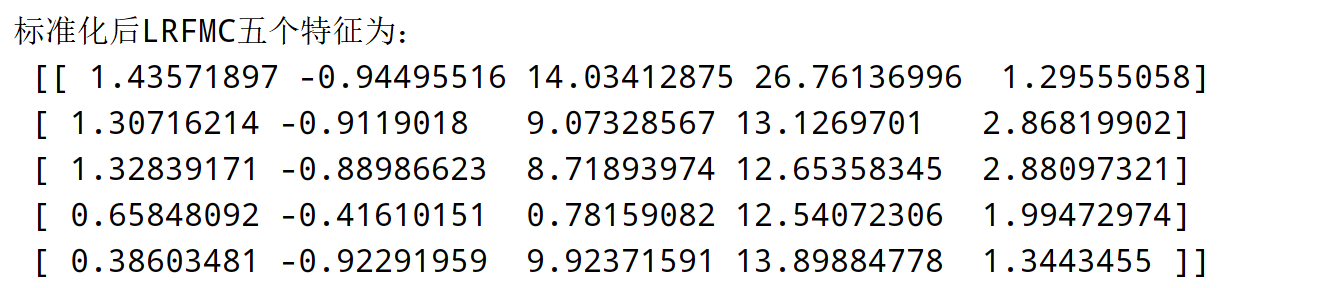
# 运行结果展示



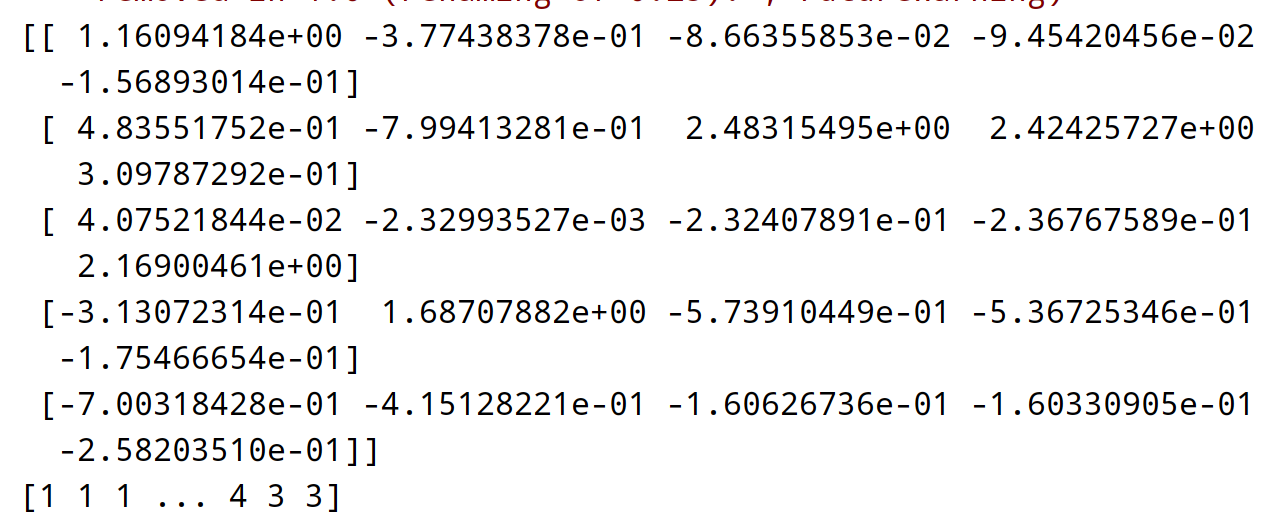
**△处理表格异常**



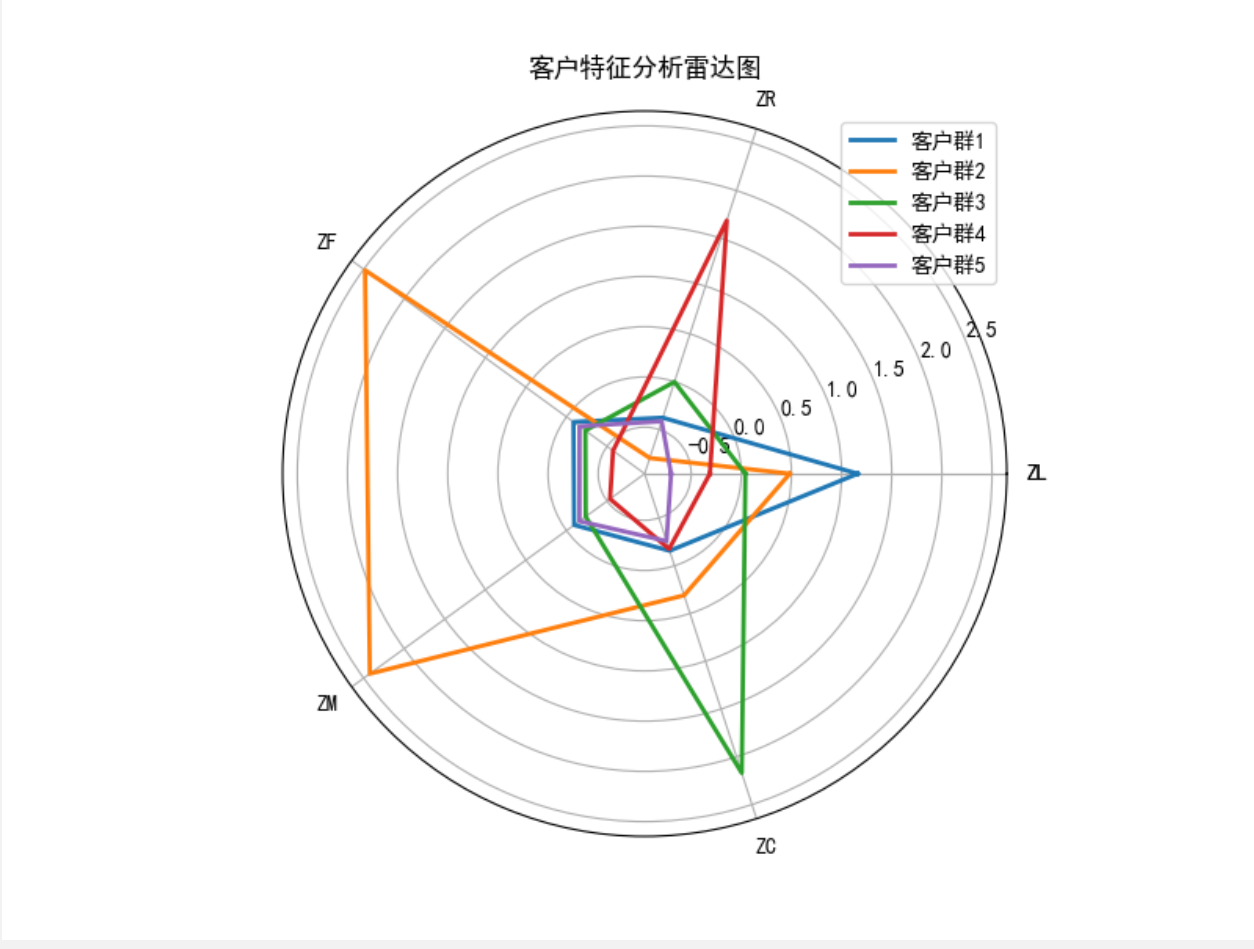
**△合并特征**



**△将特征标准化**



**△Kmeans聚类算法得到的数据**



**△客户特征雷达图**



**△聚类评价结果，可以总结出聚类数5是最佳选择。**

# 实验结果分析

结合雷达图分析：

客户1在特征L处较大，在R处较小，其他都比较适中，说明客户群1入会时间比较长，飞行频率也高，属于高价值客户群。

客户2在F和M处的值都达到了最大且R较小，说明客户2最近频繁飞行且里程较长，属于重要保持客户群。

客户3在C处的值达到了最大，其他都比较均匀且比较低，虽然这类客户目前的价值不是很高，由于他们买折扣机票较少，所以属于重要潜力客户群。

客户4在R处价值最大，其他都比较低，说明比较长时间没有乘坐过飞机且买折扣机票比较多，属于低价值客户群。

客户5在每一处价值都很小，说明是新入会的客户群，属于一般客户。

# 模型应用

根据对各个客户群进行特征分析，可以采取下面的一些营销手段和策略，为航空公司的价值客户群管理提供参考。

8.1.会员的升级与保级

航空公司的会员可以分为白金卡会员、金卡会员、银卡会员、普通卡会员等，其中非普通卡会员可以统称为航空公司的精英会员。虽然各个航空公司都有自己的特点和规定，但会员制的管理方法大同小异。成为精英会员一般都要求在一定时间内（如一年）积累一定的飞行里程或航段，达到这种要求后就会在有效期内（通常为两年）成为精英会员，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精英会员，然后对该客户进行相应的升级或降级。

然而，由于许多客户并没有意识到或根本不了解会员升级或保级的时间与要求（相关的文件说明往往复杂且不易理解），经常在评价期过后才发现自己其实只差一点就可以实现升级或保级，却错过了机会，使之前的里程积累白白损失。同时，这种认知还可能导致客户的不满，甚至放弃在本公司的消费。

因此，航空公司可以在对会员升级或保级进行评价的时间点之前，对那些接近但尚未达到要求的高价值客户进行适当提醒甚至采取一些促销活动，刺激他们通过消费达到相应标准。这样既可以使航空公司获得收益，又提高了客户的满意度，增加了航空公司的精英会员。

8.2.首次兑换

航空公司常旅客计划中最吸引客户的内容就是客户可以通过消费积累的里程来兑换免票或免费升舱等。每个航空公司都有一个首次兑换标准，当客户的里程或航段积累到一定程度时才可以实现第一次兑换，这个标准会高于正常的里程兑换标准。但是很多航空公司的里程积累随着时间的推移进行一定的削减，例如，有的航空公司会在年末对该年积累的里程进行折半处理。这样会导致许多不了解情况的会员白白损失自己好不容易积累的里程，甚至总是难以实现首次兑换。同时，这也会引起客户的不满或流失。对此可以采取的措施是：从数据库中提取出接近但尚未达到首次兑换标准的会员信息，对他们进行适当提醒或采取一些促销活动，使他们通过消费达到兑换标准。一旦实现了首次兑换，客户在本公司进行再次消费兑换就比在其他公司进行首次兑换要容易许多，等于在一定程度上提高了转移的成本。另外，在一些特殊的时间点（如里程折半的时间点）之前可以给客户一些提醒，这样可以增加客户的满意度。

8.3.交叉销售

通过发行“联名卡”等与非航空类企业的合作，使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，提高他们的忠诚度。例如，可以查看重要客户在非航空类合作伙伴处的里程积累情况，找出他们习惯的里程积累方式（是否经常在合作伙伴处消费、更喜欢消费哪些类型的合作伙伴的产品），为他们制定相应的促销策略。

在客户识别期和发展期可以为客户关系打下基石，但是这两个时期带来的客户关系是短暂的、不稳定的。企业要获取长期的利润，必须具有稳定的、高质量的客户。留住客户对于企业来说是至关重要的，不仅是因为争取一个新客户的成本远远高于维持老客户的成本，更重要的是客户流失会造成公司收益的直接损失。因此，在这一时期，航空公司应该努力维系客户关系水平，使之处于较高的水准，使生命周期内公司与客户的互动价值最大化，并使这样的高水平尽可能保持下去。对于这一阶段的客户，主要应该通过提供优质的服务产品和提高服务水平来提高客户的满意度。通过对常旅客数据库的数据分析，进行客户细分，可以获得重要保持客户的名单。航空公司应该优先将资源投放到他们身上，对他们进行差异化管理和一对一营销，提高这类客户的忠诚度与满意度，尽可能延长这类客户的高水平消费

# 心得体会

这次python实验给我的感觉是更偏重数据分析，编程时只需要补全那些算法，算是一个机器学习的简单应用。在处理繁杂的数据时，使用肉眼将其归类是非常麻烦且不精确的，而kmeans算法能自动将其以簇的形式进行划分归类，极大的提高了效率，并且有封装的比较好的函数包，也让我感受到了python这个语言的便捷性，它可以非专业人员大大提高办公效率。当然，作为计算机专业的学生，我们更要关注算法的底层实现原理，进入文档就能看出算法设计的精妙性。这次实验让我初步走进了机器学习，也没有说的那么高大上，这与我们的生活息息相关，计算机与我们的生活也是更加密不可分。