航空公司客户价值分析报告

陈寒冰 19213147、田瑶19213224、涂庆19213225

## 问题描述

### 背景与目标

随着人们生活水平的不断提高，出行方便变成了大家的关注点，而大多数人为了便捷都选择了坐飞机出行。面对激烈的竞争，航空公司推出了不同的更优惠的营销方式吸引更多客户，国内某航空公司面临着旅客流逝，竞争力下降和航空资源未充分利用等经营危机通过建立合理的客户价值评估模型，对客户群体进行分群，分析比较不同的客户群的客户价值，并制定相应的营销策略，对不同的客户群体提出个性化的服务是必须和有效的。

本文的实验目的就是对这些客户进行价值分析，从而促进航空业的发展。

本次试验主要实现以下两个目标：

1.借助航空公司客户数据，对客户进行分类。

2. 对不同的客户进行特征分析，比较不同类客户的客户价值。

### 数据集介绍

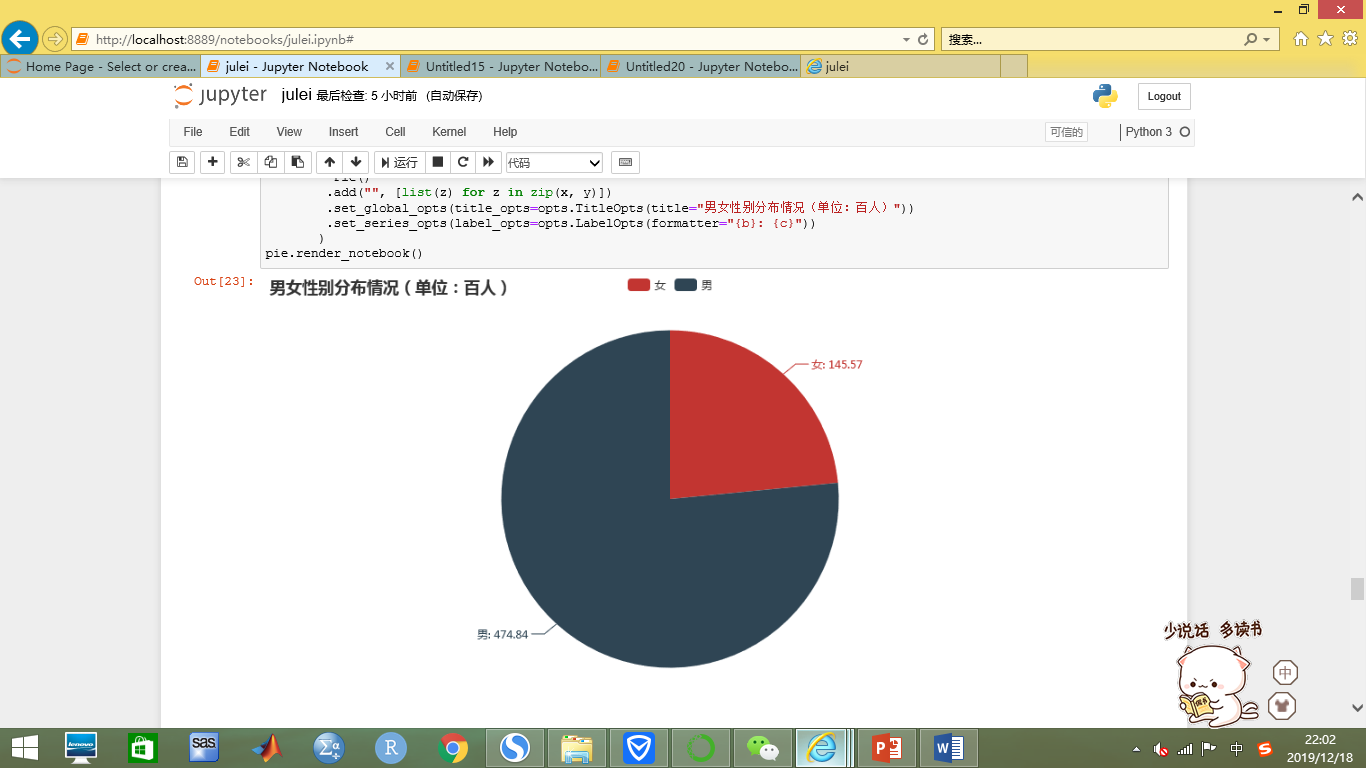
属性特征表

|  |  |
| --- | --- |
| 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 | 工作地城市 |
| WORK\_PROVINCE | 工作地所在省份 |
| WORK\_COUNTRY | 工作地所在国家 |
| age | 年龄 |
| LOAD\_TIME | 观测窗口的结束时间 |
| FLIGHT\_COUNT | 飞行次数 |
| BP\_SUM | 观测窗口总基本积分 |
| EP\_SUM\_YR\_1 | 第一年精英资格积分 |
| EP\_SUM\_YR\_2 | 第二年精英资格积分 |
| SUM\_YR\_1 | 第一年总票价 |
| SUM\_YR\_2 | 第二年总票价 |
| SEG\_KM\_SUM | 观测窗口总飞行公里数 |
| WEIGHTED\_SEG\_KM | 观测窗口总加权飞行公里数（Σ舱位折扣×航段距离） |
| LAST\_FLIGHT\_DATE | 末次飞行日期 |
| AVG\_FLIGHT\_COUNT | 观测窗口季度平均飞行次数 |
| AVG\_BP\_SUM | 观测窗口季度平均基本积分累积 |
| BEGIN\_TO\_FIRST | 观察窗口内第一次乘机时间至MAX（观察窗口始端，入会时间）时长 |
| LAST\_TO\_END | 最后一次乘机时间至观察窗口末端时长 |
| AVG\_INTERVAL | 平均乘机时间间隔 |
| MAX\_INTERVAL | 观察窗口内最大乘机间隔 |
| ADD\_POINTS\_SUM\_YR\_1 | 观测窗口中第1年其他积分（合作伙伴、促销、外航转入等） |
| ADD\_POINTS\_SUM\_YR\_2 | 观测窗口中第2年其他积分（合作伙伴、促销、外航转入等） |
| EXCHANGE\_COUNT | 积分兑换次数 |
| avg\_discount | 平均折扣率 |
| P1Y\_Flight\_Count | 第1年乘机次数 |
| L1Y\_Flight\_Count | 第2年乘机次数 |
| P1Y\_BP\_SUM | 第1年里程积分 |
| L1Y\_BP\_SUM | 第2年里程积分 |
| EP\_SUM | 观测窗口总精英积分 |
| ADD\_Point\_SUM | 观测窗口中其他积分（合作伙伴、促销、外航转入等） |
| Eli\_Add\_Point\_Sum | 非乘机积分总和 |
| L1Y\_ELi\_Add\_Points | 第2年非乘机积分总和 |
| Points\_Sum | 总累计积分 |
| L1Y\_Points\_Sum | 第2年观测窗口总累计积分 |
| Ration\_L1Y\_Flight\_Count | 第2年的乘机次数比率 |
| Ration\_P1Y\_Flight\_Count | 第1年的乘机次数比率 |
| Ration\_P1Y\_BPS | 第1年里程积分占最近两年积分比例 |
| Ration\_L1Y\_BPS | 第2年里程积分占最近两年积分比例 |
| Point\_NotFlight | 非乘机的积分变动次数 |

## 数据预处理

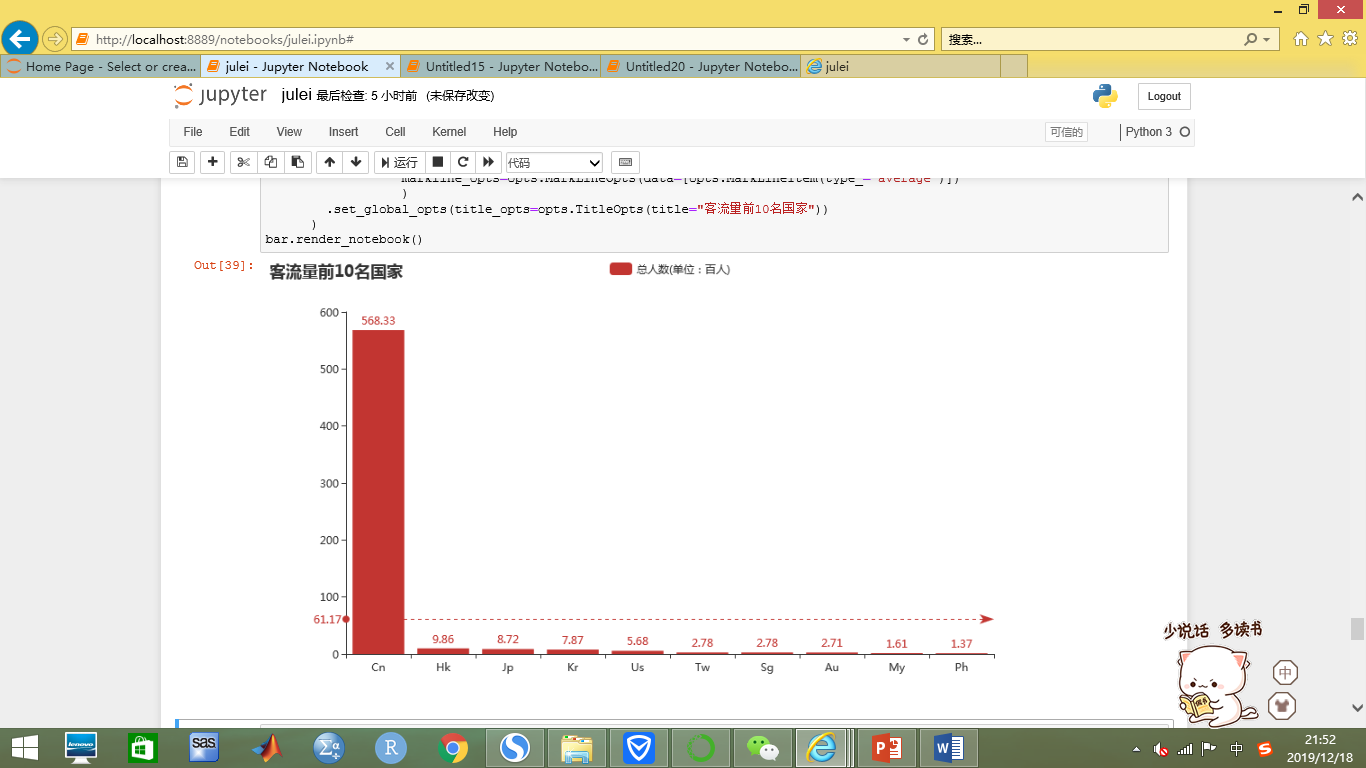
### 描述性统计

**男女性别分布情况图**



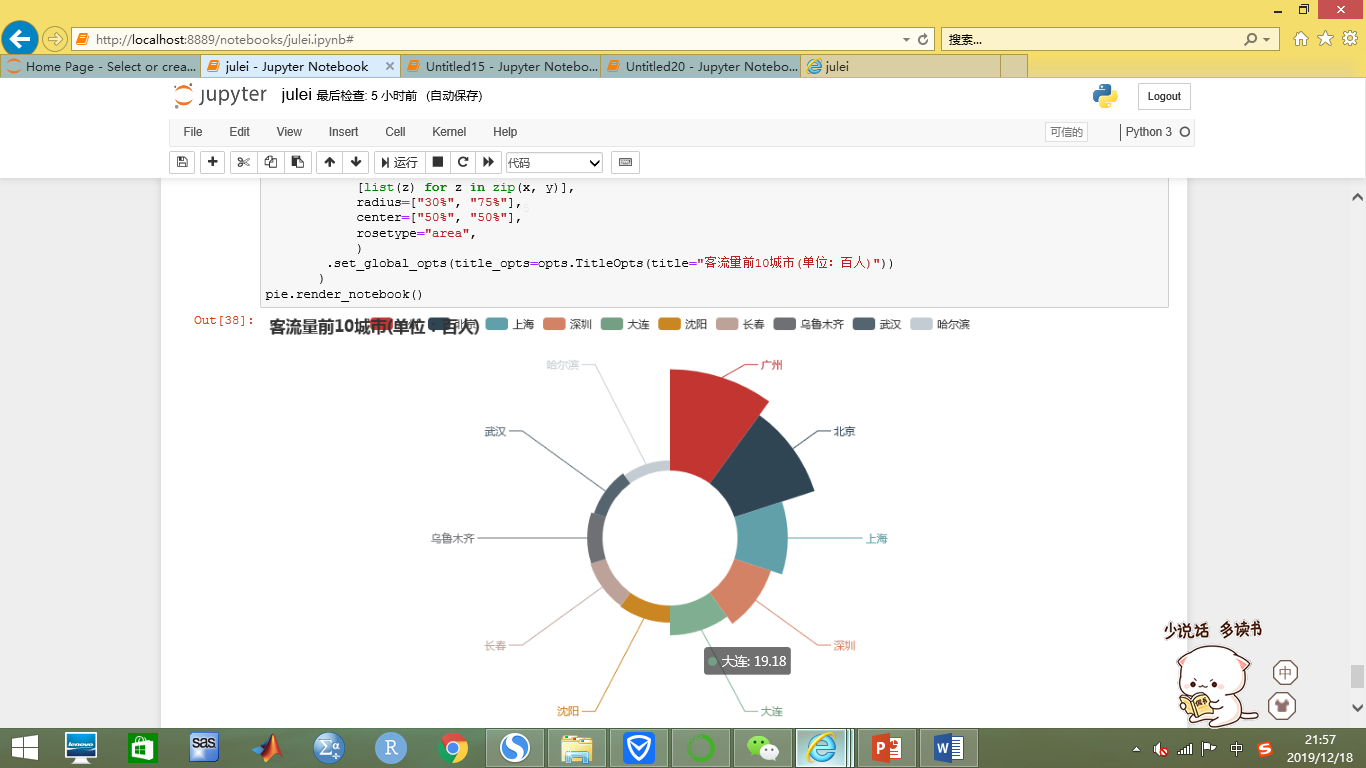
从性别来看，男性乘客人数大约是女性乘客的3倍。

**客流量前10国家**



中国人数遥遥领先，说明该航空公司的国内航班较多，极有可能是国内的公司。

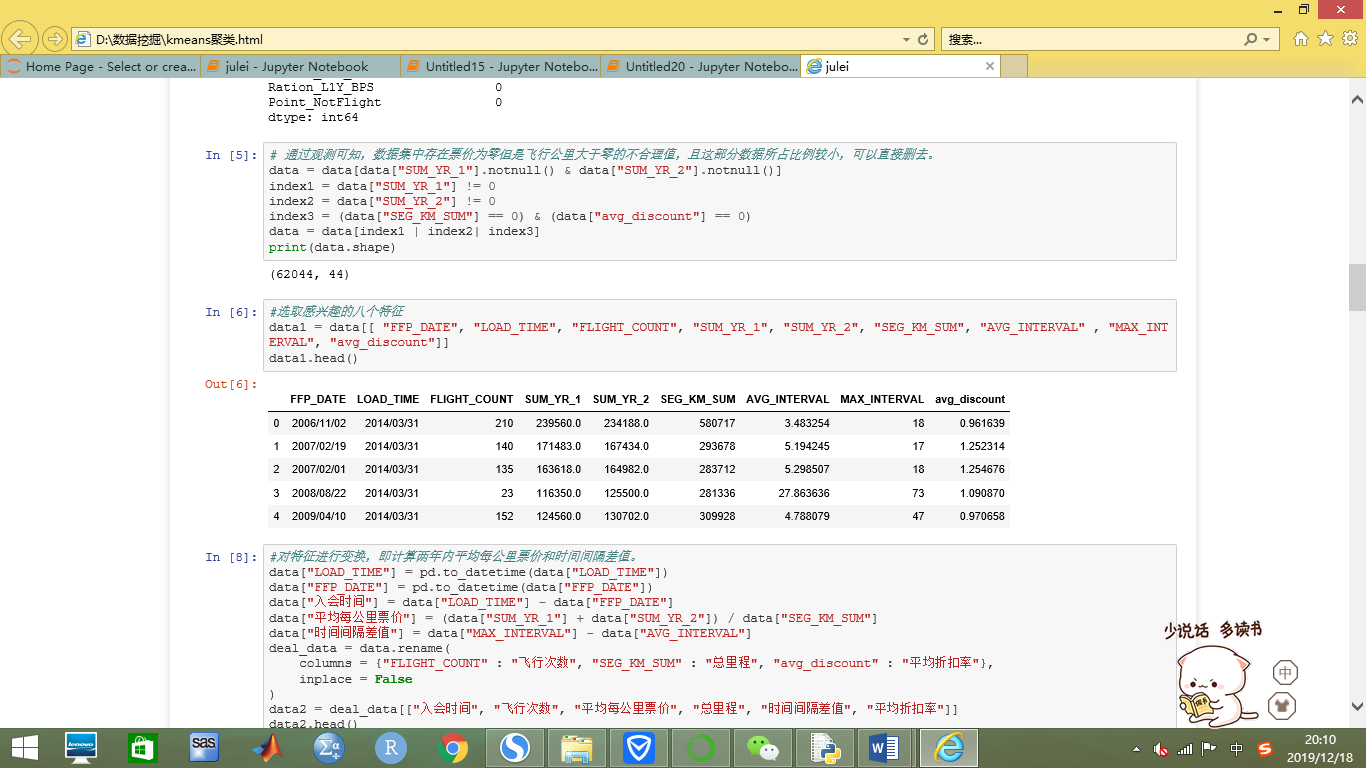
**客流量前10城市（单位：百人）**



客流量前10的城市都是国内城市，同样佐证了航空公司是国内公司，且客流量上发达城市北上广深领先中国人数遥遥领先。

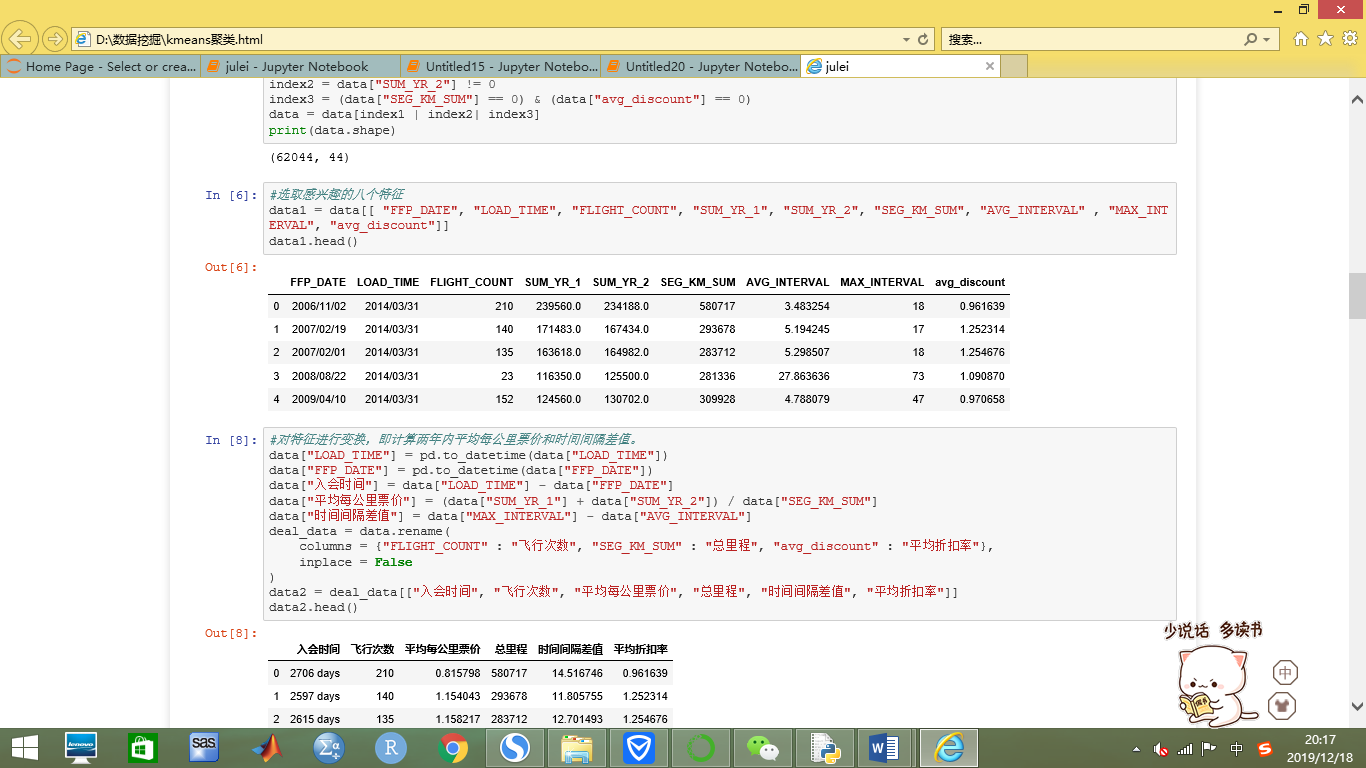
### 数据清洗

通过观测，数据集中存在票价为零但是飞行公里大于零的不合理值，且这部分数据所占比例较小，可以直接删去。删除后剩余的样本值是62044个，可见异常样本的比例不足1.5%，因此不会对分析结果产生较大的影响。



### 属性归约

原始数据集的特征属性太多，共有44个，而且各属性不具有降维的特征，故这里只选取几个对航空公司来说比较有价值的特征进行分析，最终一共选取了8个特征，选取的特征是：第一年总票价（SUM\_YR\_1）、第二年总票价（SUM\_YR\_2）、观测窗口总飞行公里数（SEG\_KM\_SUM）、飞行次数（FLIGHT\_COUNT）、平均乘机时间间隔（AVG\_INTERVAL）、观察窗口内最大乘机间隔（MAX\_INTERVAL）、入会时间（FFP\_DATE）、观测窗口的结束时间（LOAD\_TIME）、平均折扣率（avg\_discount）这八个特征。

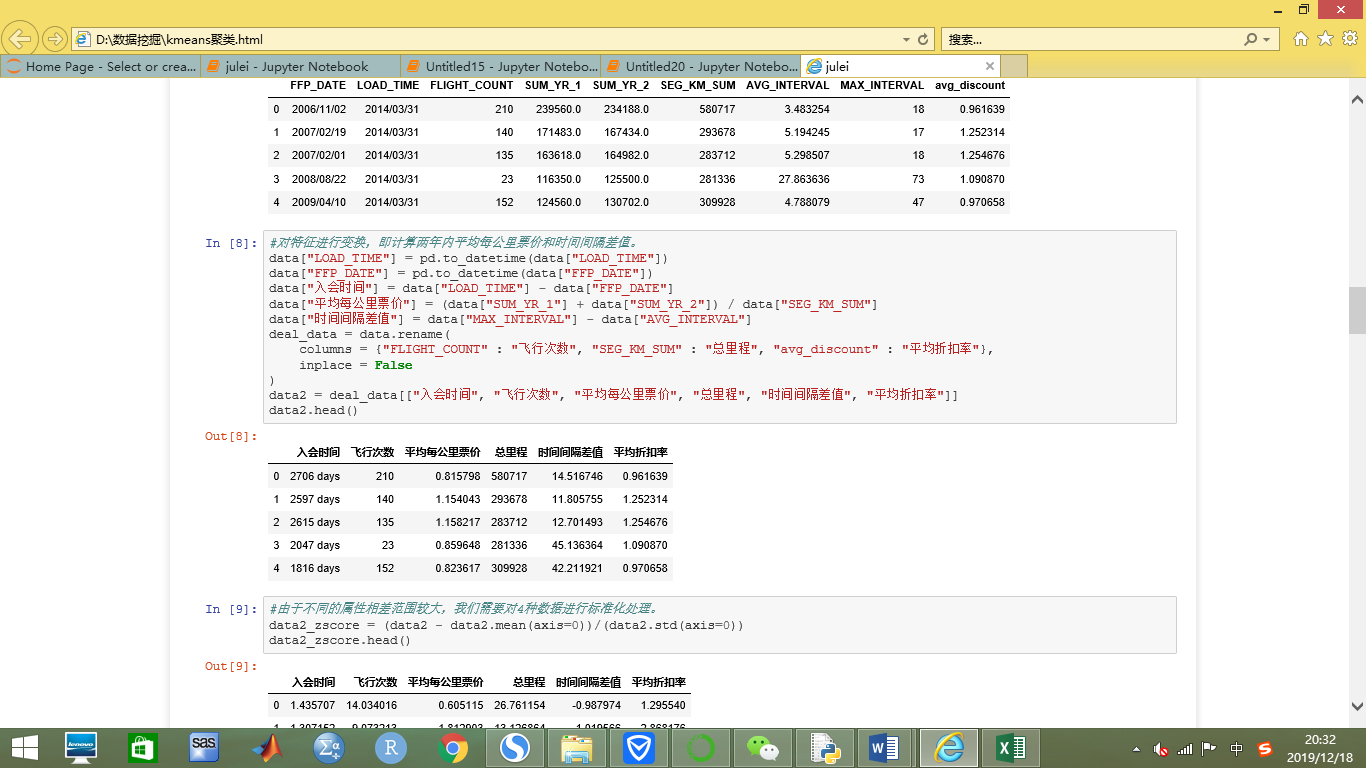


特征选择的理由：

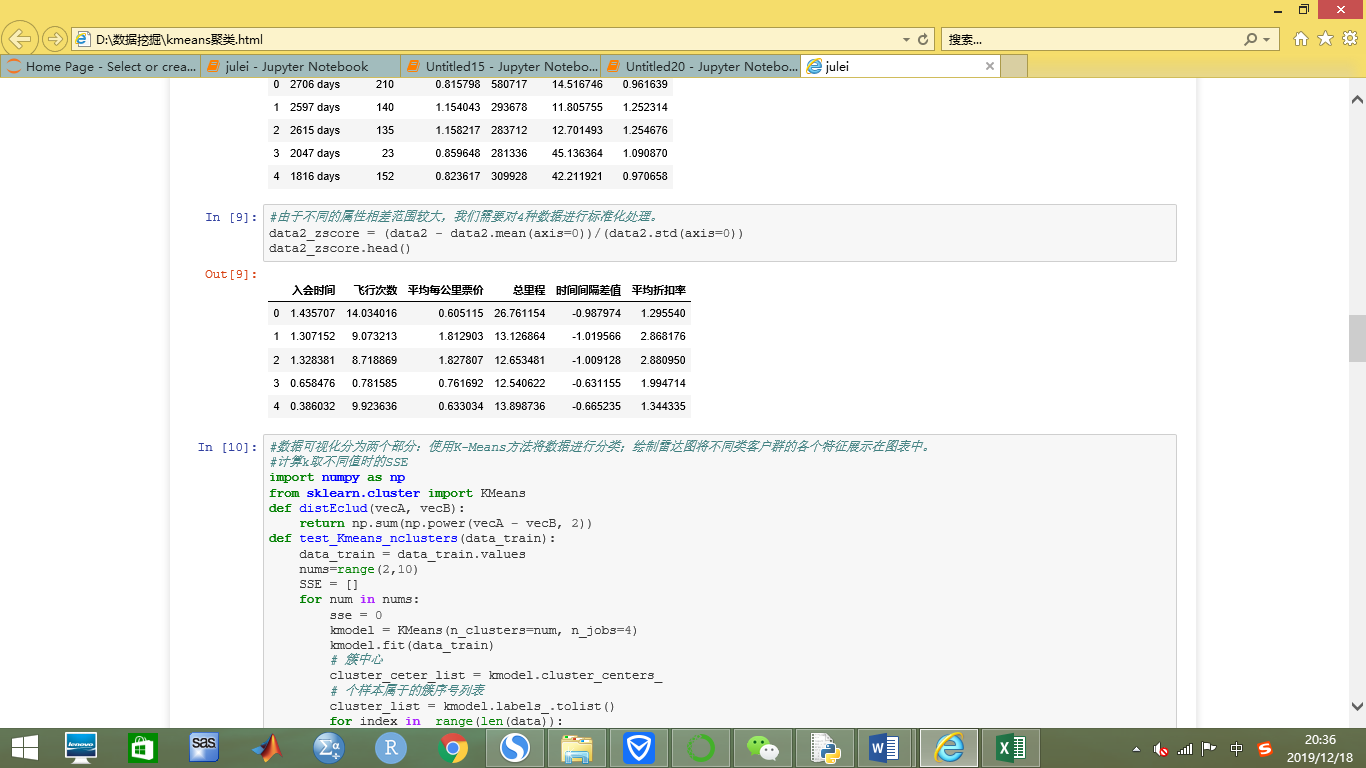
1. 选取的特征是第一年总票价、第二年总票价、观测窗口总飞行公里数是要计算平均飞行每公里的票价，因为对于航空公司来说并不是票价越高，飞行公里数越长越能创造利润，相反而是那些近距离的高等舱的客户创造更大的利益；
2. 总飞行公里数、飞行次数也都是评价一个客户价值的重要的指标；
3. 入会时间可以看出客户是不是老用户及其忠诚度；
4. 通过平均乘机时间间隔、观察窗口内最大乘机间隔可以判断客户的乘机频率是不是固定；
5. 平均折扣率可以反映出客户给公司带来的利益，毕竟来说越是高价值的客户享用的折扣率越高。

### 特征变换与数据标准化

为便于观察，进行特征变换，计算两年内平均每公里票价和时间间隔差值，并对特征重新命名，结果如下：



由于不同的属性相差范围较大，我们需要对4种数据进行标准化处理，结果如下：



1. 聚类分析
2. k-means 聚类算法

#### 1.1 k-means 聚类算法原理

K-means算法是最常用的一种聚类算法。算法的输入为一个样本集（或者称为点集），通过该算法可以将样本进行聚类，具有相似特征的样本聚为一类。

针对每个点，计算这个点距离所有中心点最近的那个中心点，然后将这个点归为这个中心点代表的簇。一次迭代结束之后，针对每个簇类，重新计算中心点，然后针对每个点，重新寻找距离自己最近的中心点。如此循环，直到前后两次迭代的簇类没有变化。

#### 1.2 k-means聚类算法步骤

实质是EM算法的模型优化过程，具体步骤如下：

1）随机选择k个样本作为初始簇类的均值向量；

2）将每个样本数据集划分离它距离最近的簇；

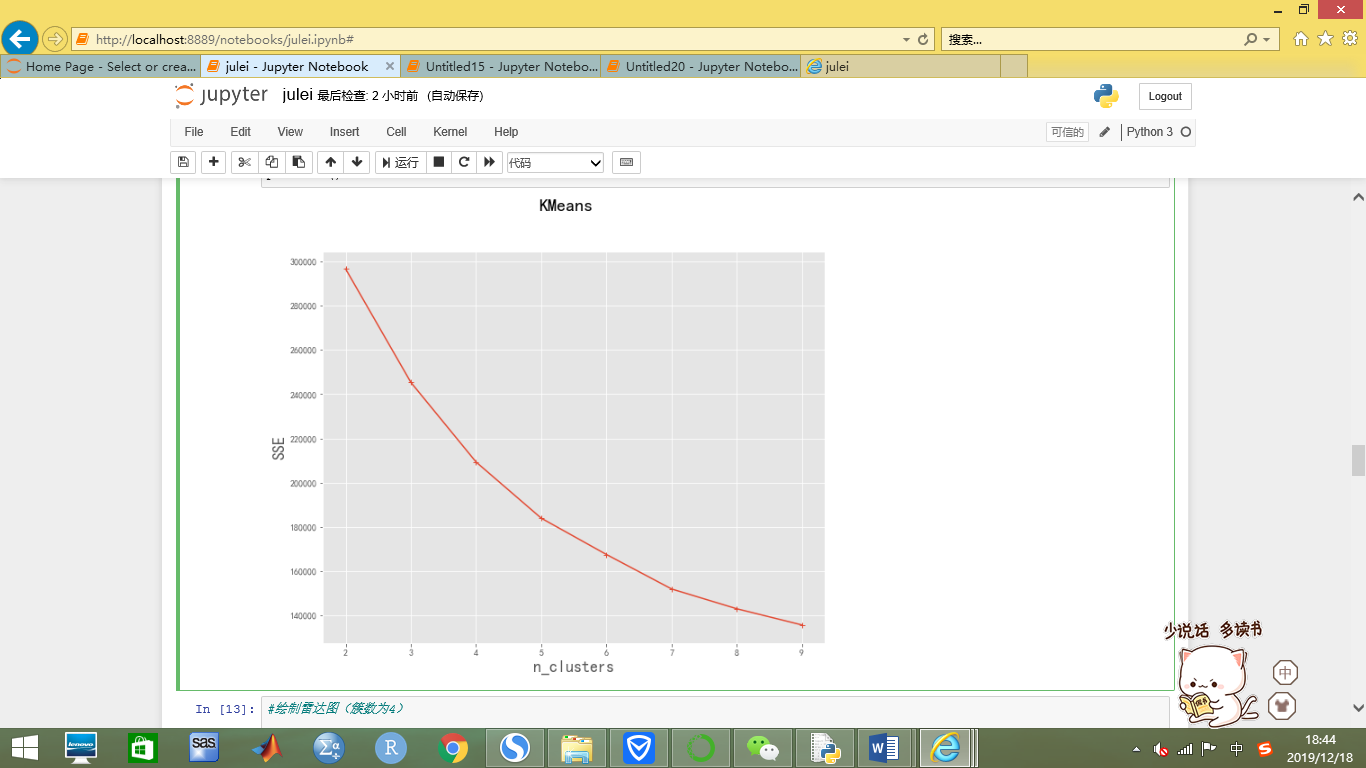
3）根据每个样本所属的簇，更新簇类的均值向量；

4）重复（2）（3）步，当达到设置的迭代次数或簇类的均值向量不再改变时，模型构建完成，输出聚类算法结果。

#### 1.3 计算SSE判断分类簇数k

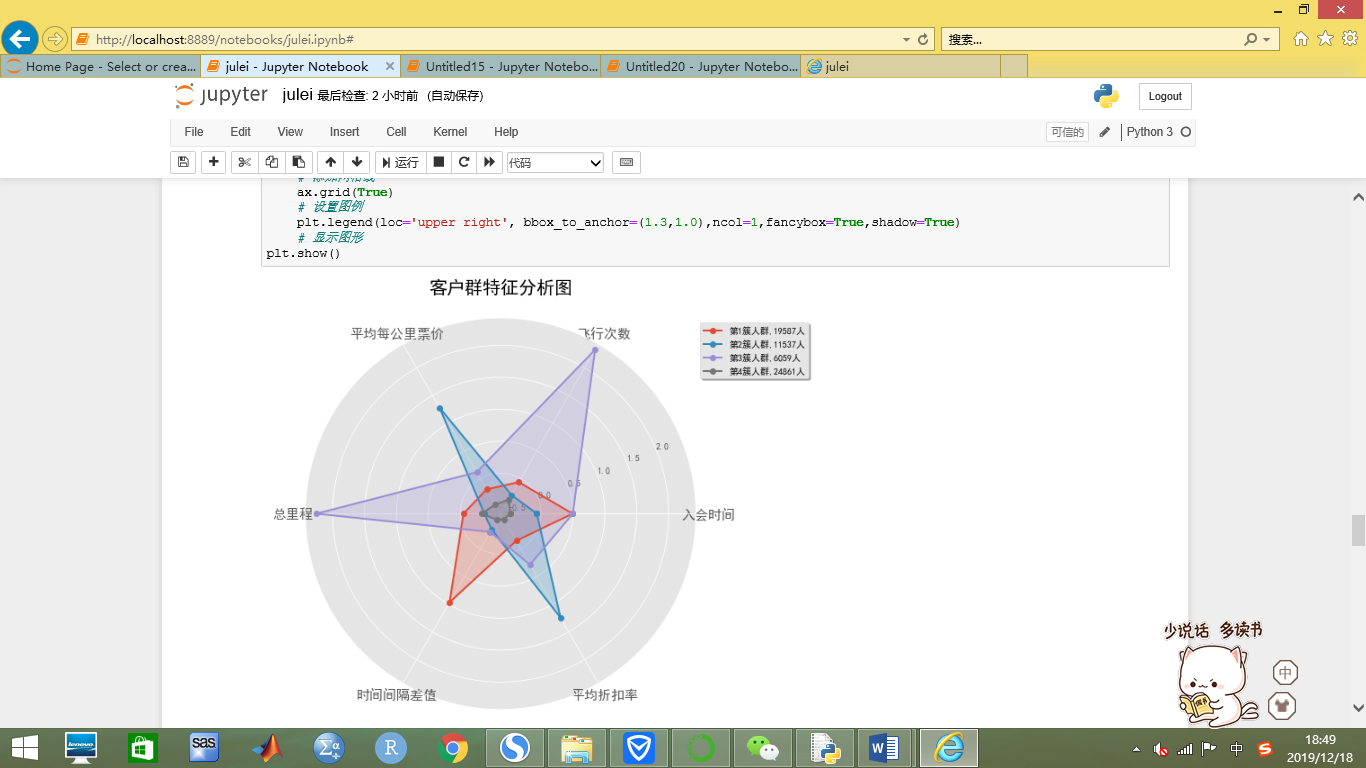
因为是无监督的聚类分析问题，对于K-Means方法，k的取值是不确定的，并且不存在绝对正确的值，需要进行研究试探。此处我们采用计算SSE的方法，尝试找到最优的k值。

画出SSE随聚类数k的变化情况图像，发现并没有出现斜率的大幅变动，即没有所谓的“肘”点出现，SSE是随k值的增大逐渐减小的，这里选取当k分别取4， 5， 6时进行分类，看能不能通过分析结果来反向选取更合适的值。

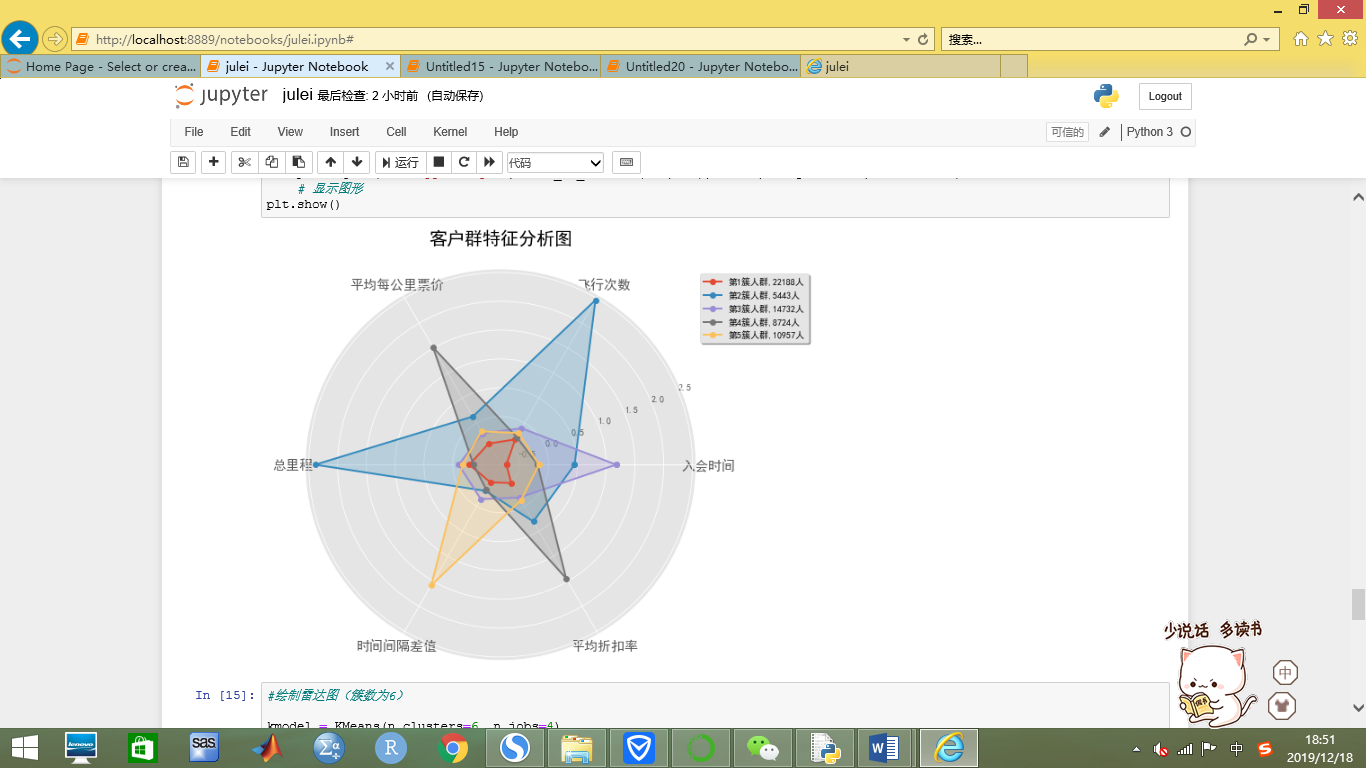


#### 1.4 聚类结果——雷达图展示

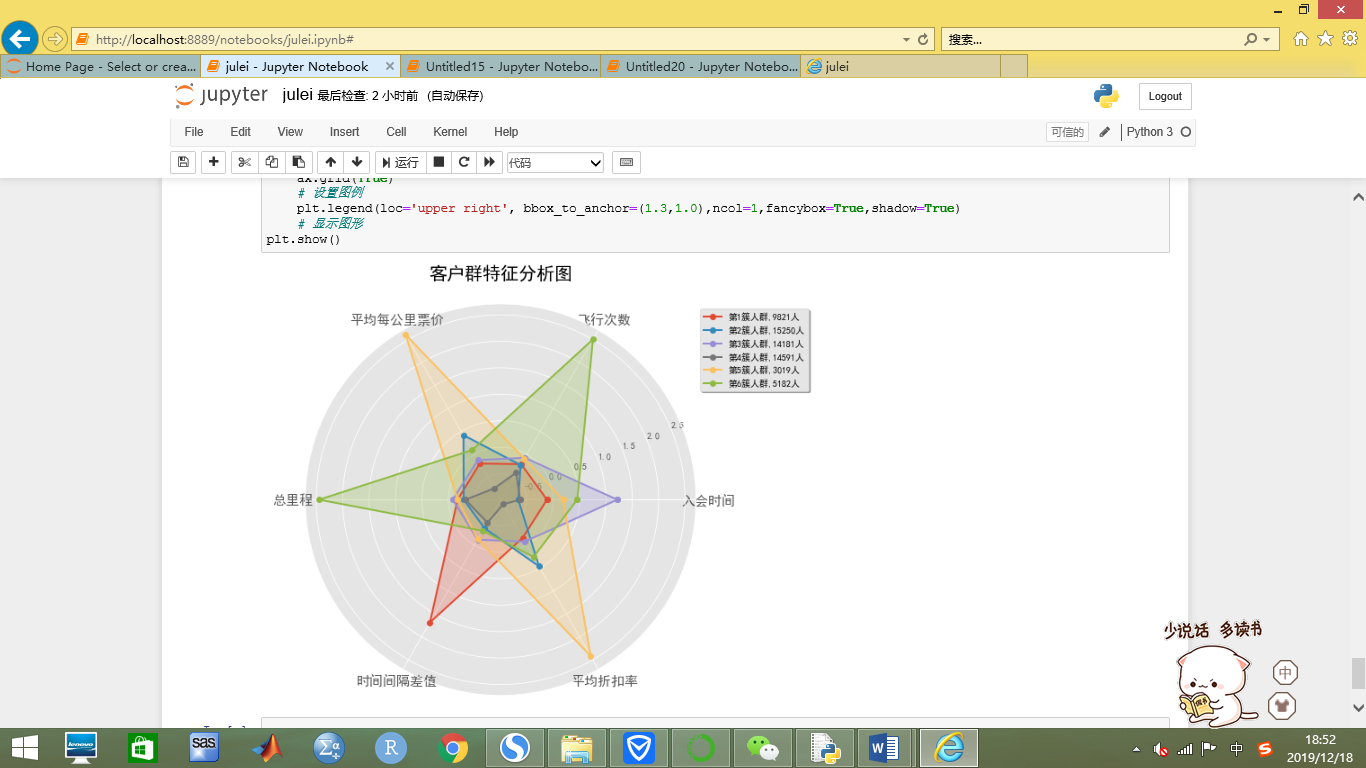
**分类簇数为4的雷达图**



**分类簇数为5的雷达图**



**分类簇数为6的雷达图**



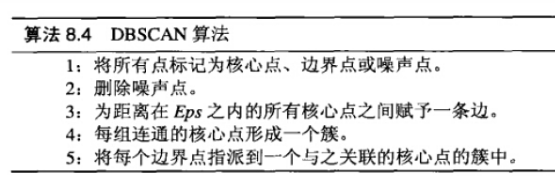
通过观察对比上面三张图可以看出：当k取值为4时，每个人群包含的信息比较复杂，且特征不明显；当k取值5时，分析的结果比较合理，分出的五种类型人群都有自己的特点又不相互重复；当k取值6时，各种人群也都有自己的特点，但是第2簇人群完全包含在了第5簇人群特征中，有点冗余，所以取k值为5较为合理。

1. BDSCAN算法

1.1算法思想与原理

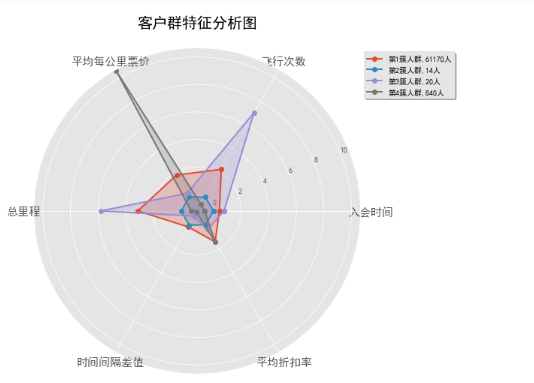
DBSCAN（Density-Based Spatial Clustering of Applications with Noise，具有噪声的基于密度的聚类方法）是一种很典型的密度聚类算法。与只适用于凸样本集的K-Means聚类相比，DBSCAN既可以适用于凸样本集，也可以适用于非凸样本集。  
 DBSCAN一般假定类别可以通过样本分布的紧密程度决定。同一类别的样本，他们之间的紧密相连的，也就是说，在该类别任意样本周围不远处一定有同类别的样本存在。通过将紧密相连的样本划为一类，这样就得到了一个聚类类别。通过将所有各组紧密相连的样本划为各个不同的类别，则我们就得到了最终的所有聚类类别结果。

1.2算法过程

**DBSCAN最大的不同就是不需要输入类别数k**，当然它最大的优势是可以发现任意形状的聚类簇，而不是像K-Means，一般仅仅使用于凸的样本集聚类。同时它在聚类的同时还可以找出异常点，这点和BIRCH算法类似。算法的过程如下：  


1.3 算法结果

DBSCAN算法自动选择了将客户分为4类，将聚类结果可视化如下：



分类簇数由BDCSAN算法自身确定，最终结果为4，同时可以对比K-Means聚类可得，该算法聚类将大量的客户聚集在第1簇中，同时每一类的信息区分度不是很大，重合度很高，因此聚类效果不佳。

1. 总结

综上，我们最终选择K-Means聚类下k=5时的聚类结果为最佳**。**

得到最终聚类而总结得到的客户价值排名表：

|  |  |  |  |
| --- | --- | --- | --- |
| 客户群体 | 群体人数 | 价值排名 | 价值种类 |
| 第一簇人群 | 22188人 | 4 | 一般价值客户 |
| 第二簇人群 | 5443人 | 1 | 重要-保持客户 |
| 第三簇人群 | 14732人 | 3 | 重要-挽留客户 |
| 第四簇人群 | 8724人 | 2 | 重要-发展客户 |
| 第五簇人群 | 10957人 | 5 | 较低价值客户 |

将所有的客户分成5类，再根据选取的特征进一步分析可以得到以下结论：

1. **一般价值客户：**第一簇人群，共有22188人，在各方面的数据都比较低，这类客户不常选用飞机作为交通工具，属于一般价值用户；

**建议：**充分利用航空资源，适时放出特价优惠机票，增加对这类客户的吸引力。

1. **重要-保持客户**：第二簇人群，共有5443人，总里程和飞行次数都是最多的，而且入会时间较长，说明他们经常出行，可能是商务出差人员，且是本航空公司的忠实用户，此外，他们的时间间隔差值小，出行频繁，是航空公司的理想客户，对于这类客户，航空公司应该重点保持；

**建议：**针对这一类客户群体，航空公司尽可能将资源放到他们身上，制定一对一的精准营销，提高客户的满意度，使得这类人群的消费周期更长一些，维持他们的消费忠诚度。

1. **重要-挽留客户：**第三簇人群，共有14732人，这群人最明显的特点就是入会的时间较长，应该属于老客户，但是他们的平均折扣率、总里程和总次数都不高，极有可能是流失的客户，对于这类客户，公司应该再争取一下，让老客户回头；

**建议：**航空公司因根据这些客户最近的消费时间，次数，推测客户消费异动情况，制定相应的策略，尽可能的挽回这类客户，提高这类客户的价值。

1. **重要-发展客户：**第四簇人群，共有8724人，这类人群平均每公里票价和平均折扣率都是最高的，应该属于乘坐高等舱的人员，应该重点保持的客户，但是他们的飞行次数、总里程都不高，也是公司需要重点发展的客户，另外应该积极采取相关的优惠政策是他们的乘坐次数增加；

**建议：**航空公司须制定营销策略，比如客户如果选择本航空公司达到一定的次数时，升级会员，附赠更多的优惠方式，努力吸引这类客户在本航空公司消费，提供高质量的服务，增加这类客户的满意度，使这类客户成为航空公司的忠实客户。

1. **较低价值客户：**第五簇人群，10957人，这类客户入会时间和平均折扣率的值都比较小，飞行里程短，飞行次数少，最大的特点是时间间隔差值最大，分析可能是“季节型客户”，一年中在某个时间段需要乘坐飞机进行旅行，其他的时间则出行的不多，或只有在航空公司出售低价机票时才会选择该航空公司，这类客户我们需要在保持的前提下，进行一定的发展；

**建议：**航空公司可以经常放出特价优惠机票，充分利用航空资源，增加这些客户对本公司的关注度。