**《大数据分析原理与实践》**

**课程设计**

|  |  |  |
| --- | --- | --- |
| **组员** | 姓名 | 李若昊 |
| 学号 | 5120180269 |
| **组员** | 姓名 |  |
| 学号 |  |
| **组员** | 姓名 |  |
| 学号 |  |
| **组员** | 姓名 |  |
| 学号 |  |

**指导老师: 彭莉娟 讲师**

**2021年6月**

**课 程 成 绩 表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **项目** | **评 分 细 则** | | | | **成 绩** |
| **案 例**  **质 量**  **（60分）** | **目标任务** | 4-5分：数据分析目标明确、任务具体 | | |  |
| 3-4分：数据分析目标基本明确、主要任务清楚 | | |
| 0-3分：数据分析目标模糊、任务单一或过于简单 | | |
| **数据预处理** | 8-10分：数据预处理方法正确，处理结果符合要求 | | |  |
| 6-8分：数据预处理方法基本正确 | | |
| 0-6分：数据预处理方法不恰当或有缺陷 | | |
| **数据探索** | 12-15分：数据探索分析方法正确，可视化美观 | | |  |
| 9-12分：数据探索分析方法基本正确 | | |
| 0-9分：数据探索分析过于简单 | | |
| **数据分析模型** | 20-25分：数据分析模型选择合理、结果正确 | | |  |
| 15-20分：数据分析模型选择基本合理，有确切结果 | | |
| 0-15分：数据分析模型选择存在缺陷，无确切结果 | | |
| **方案评估** | 4-5分：能较全面、客观、准确地评估课程设计方案 | | |  |
| 3-4分：能一定程度对课程设计方案的某些方面做出评估 | | |
| 0-3分：对课程设计方案的评估虚浮，缺乏有效性 | | |
| **文 档**  **质 量**  **（20分）** | 16-20分：文档结构完整，内容充实，重点突出，撰写规范 | | | |  |
| 12-16分：文档结构基本完整，内容较充实，规范性尚可 | | | |
| 0-12分：文档结构散乱，内容空洞浮泛，排版混乱 | | | |
| **答辩**  **（20分）** |  | |  |  | |
|  | |  |  | |
| **总分** |  | |  |  | |

**信用卡客户风险识别模型构建**

**摘要：**根据信用卡信息数据说明表，熟悉信用卡的基本业务知识，并进行异常值的识别与处理。通过查阅资料，了解信用卡模型的原理，尝试构建信用卡用户风险分析关键特征。

关键词：信用卡，风险，模型

目 录

一、数据分析目标与任务………………………………………1

二、数据预处理…………………………………………………2

2.1 数据说明…………………………………………………….2

2.2 数据清洗…………………………………………………….3

2.3数据处理……………………………………………………..4

三、数据探索……………………………………………………5

3.1 探索性分析………………………………………………….5

3.2 可视化结果………………………………………………...11

四、数据分析模型……………………………………………..12

4.1 选择模型…………………………………………………...12

4.2 模型评估…………………………………………………...13

五、方案评估…………………………………………………..15

1. **数据分析目标与任务**

目标：

1. 熟悉信用卡的基本业务知识
2. 掌握异常值的识别与处理方法
3. 掌握信用卡模型的原理
4. 构建信用卡用户风险分析关键特征
5. 掌握K-Means聚类算法的的应用
6. 掌握聚类算法结果分析的方法

任务：

1. 读取信用卡数据
2. 丢弃逾期、呆账、强制停卡、退票、拒往记录为1、瑕疵户为2的记录。
3. 丢弃呆账、强制停卡、退票为1、拒往记录为2的记录。
4. 丢弃频率为5、刷卡金额不等于1的数据。
5. 根据特征瑕疵户、逾期、呆账、强制停卡记录、退票、拒往记录，构建历史行为特征。
6. 根据特征借款余额、个人月收入、个人月开销、家庭月收入和月刷卡额，构建出经济风险情况特征
7. 根据特征职业、年龄、住家，构建出收入风险情况特征
8. 标准化历史行为、经济风险情况、收入风险情况特征
9. 构建K-Means聚类模型，聚类数为5
10. 训练K-Means聚类模型，并求出聚类中心、每类的用户数目。

编程语言：Python 3.8

开发环境：Jupyter Notebook

**二、数据预处理**

1. 数据说明

数据规模：65535 行，28 列

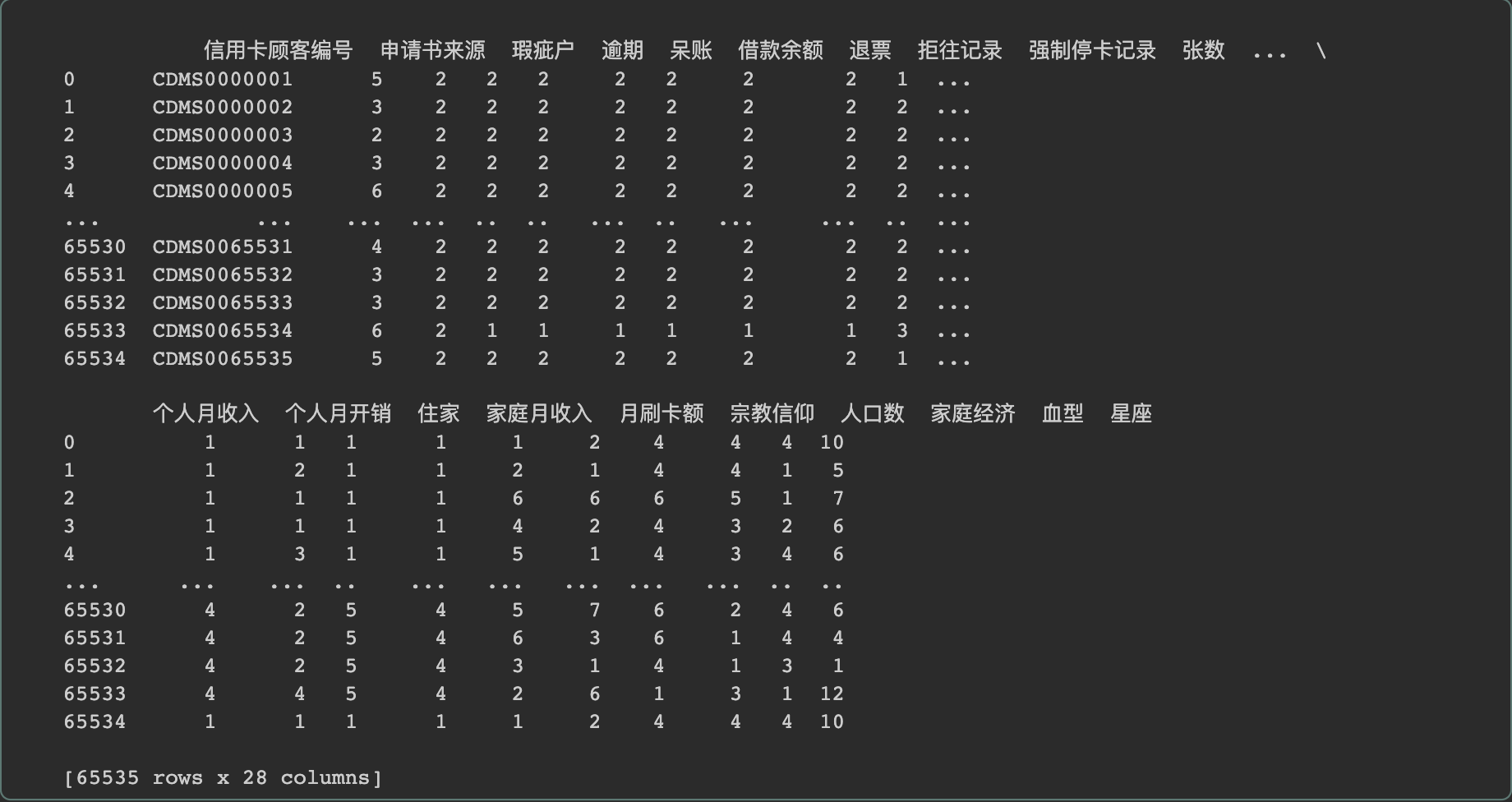
数据文件：credit\_card.csv

各字段及说明：



数据样本：

利用pandas读取CSV文件，并输出，可见：



1. 数据清洗

清洗：

1. 丢弃逾期、呆账、强制停卡、退票、拒往记录为1、瑕疵户为2的记录。



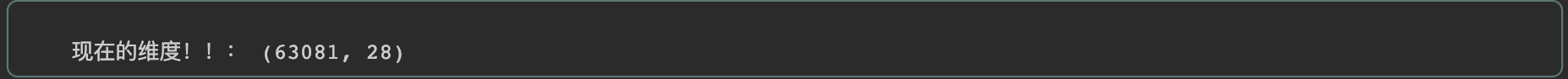
效果：这一步之后数据维度从原始的65535行缩减到63339行：



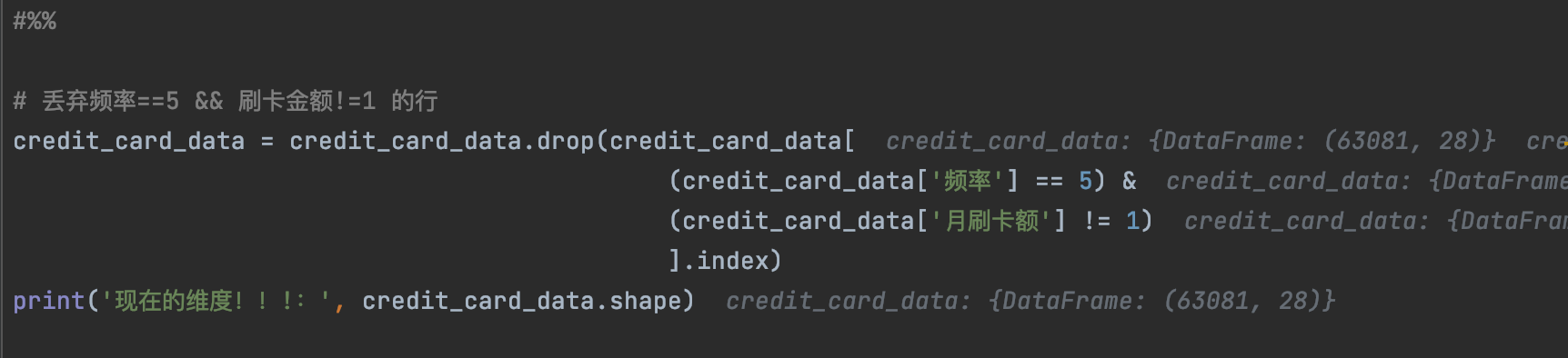
1. 丢弃呆账、强制停卡、退票为1、拒往记录为2的记录。



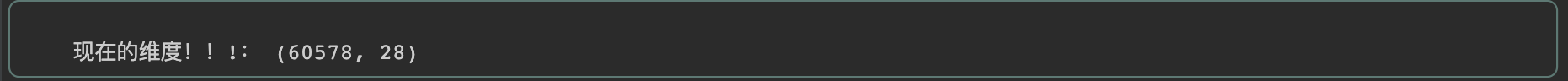
效果：这一步之后数据维度进一步缩减到63081行：



1. 弃频率为5、刷卡金额不等于1的数据。



效果：这一步之后数据维度进一步缩减到60578行：



原理是它们都是异常值。例如拿（1）来说，如果逾期、呆账、强制停卡、退票、拒往记录均为1，那这位顾客应该是瑕疵户才对；但是原纪录中瑕疵户却被标记为2: 非瑕疵户，所以这里当作异常值直接丢弃。

处理后的数据维度为(60578, 28)，相比于初始的63335行除去了数千条数据，可见总体效果还比较显著。

1. 数据处理（本项目数据预处理的难点在哪里？理由是什么？）

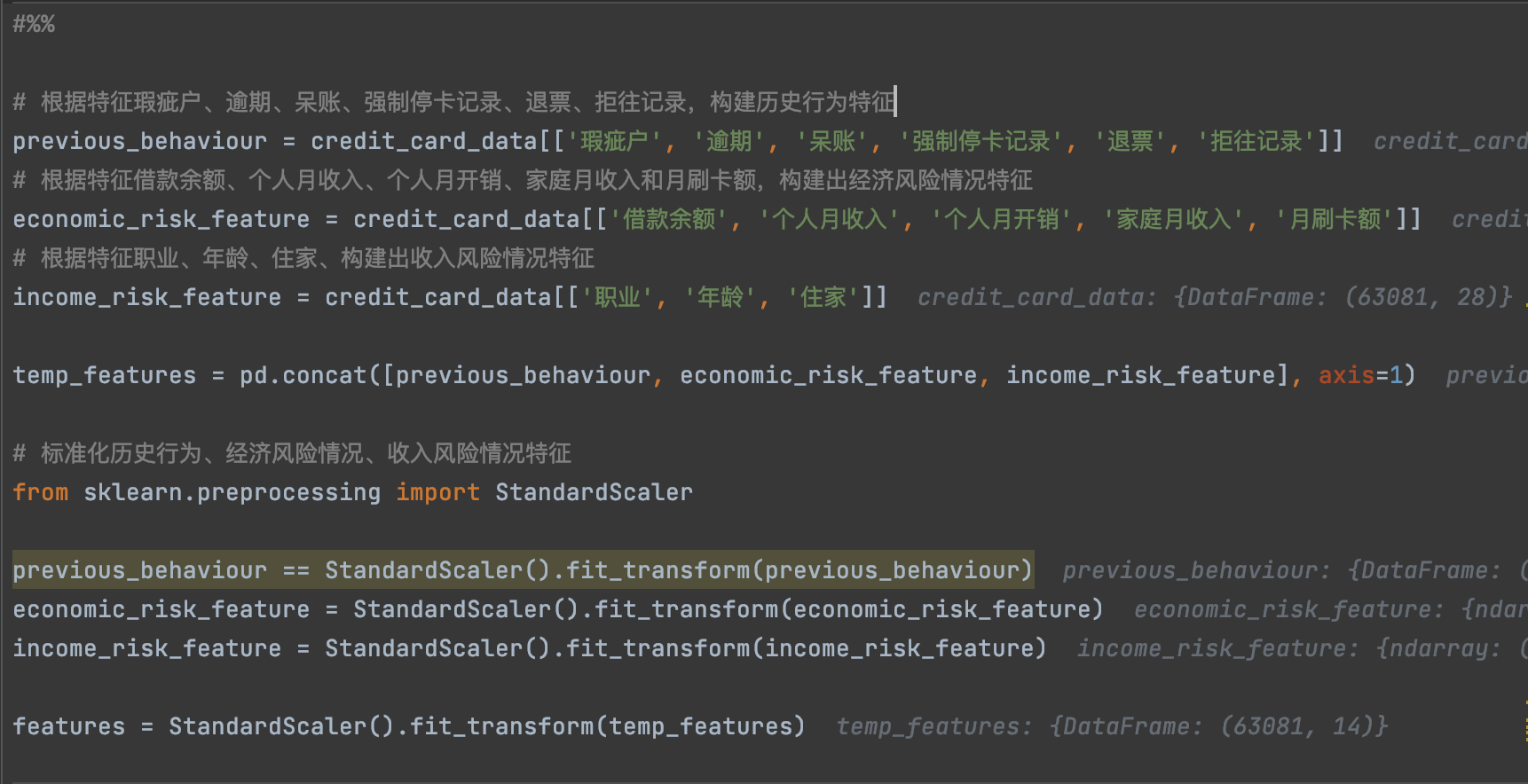
特定数据提取：

1. 根据特征瑕疵户、逾期、呆账、强制停卡记录、退票、拒往记录，构建历史行为特征。

（2）根据特征借款余额、个人月收入、个人月开销、家庭月收入和月刷卡额，构建出经济风险情况特征。

（3）根据特征职业、年龄、住家，构建出收入风险情况特征。

（4）标准化历史行为、经济风险情况、收入风险情况特征。



难点：求得三个特征之后，标准化时的操作：一开始我直接分别把三类各自标准化，这个操作确实可行，但是之后想把这三个特征连接在一起的时候就会出现数据结构上的问题；后来我尝试先连接再标准化，就没有再出现问题。

**三、数据探索分析**

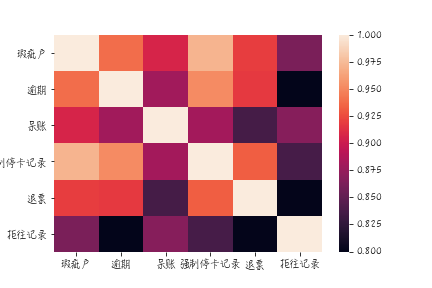
1. 结合可视化呈现，对数据进行探索性分析

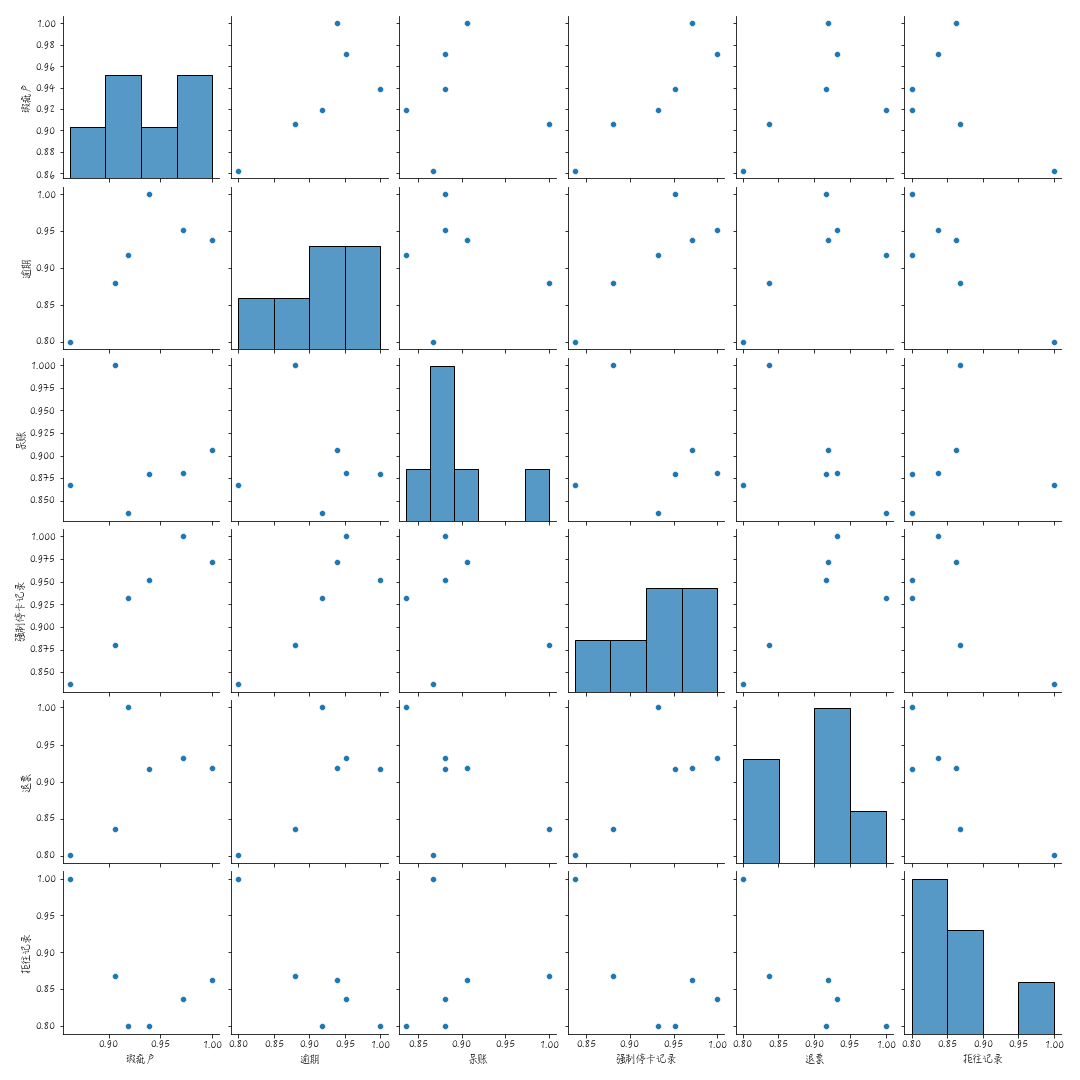
从数据相关性进行探索性分析：

1. 由常识得，历史行为特征应与瑕疵户、逾期、呆账、强制停卡记录、退票、拒往记录，构建有关，进行验证：



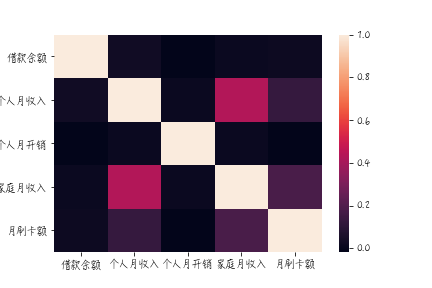
可视化呈现：

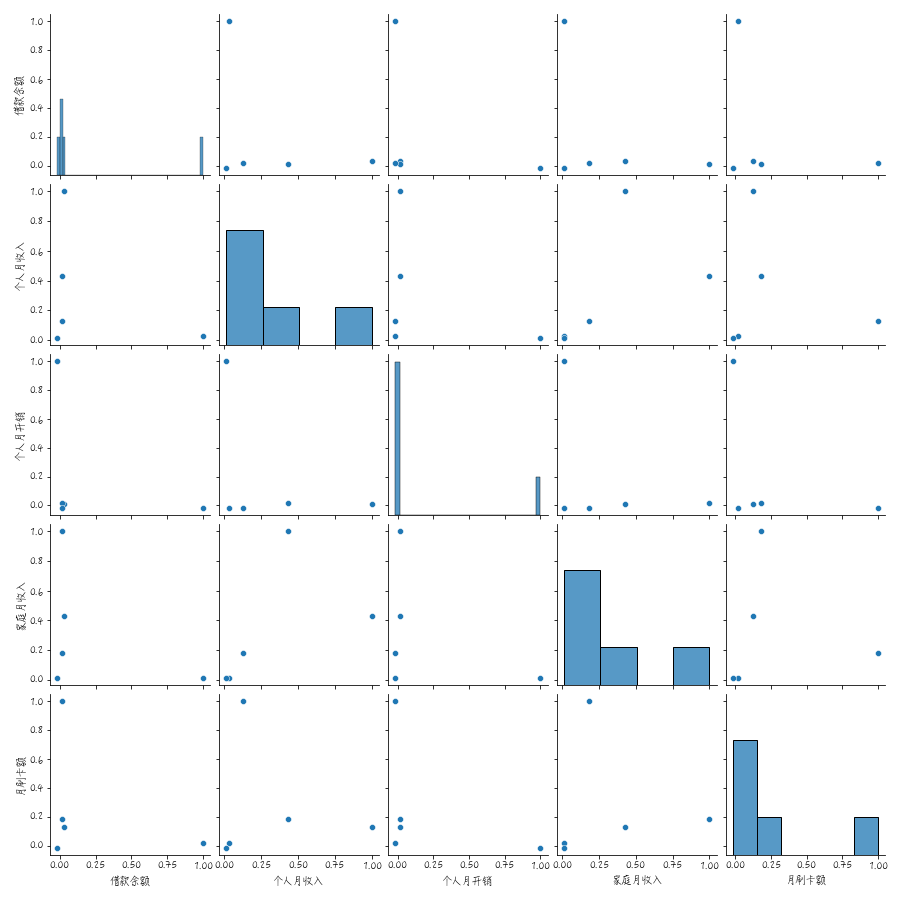




1. 由常识得，经济风险情况特征应与借款余额、个人月收入、个人月开销、家庭月收入和月刷卡额有关，验证得：

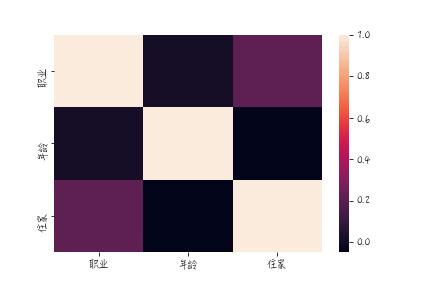


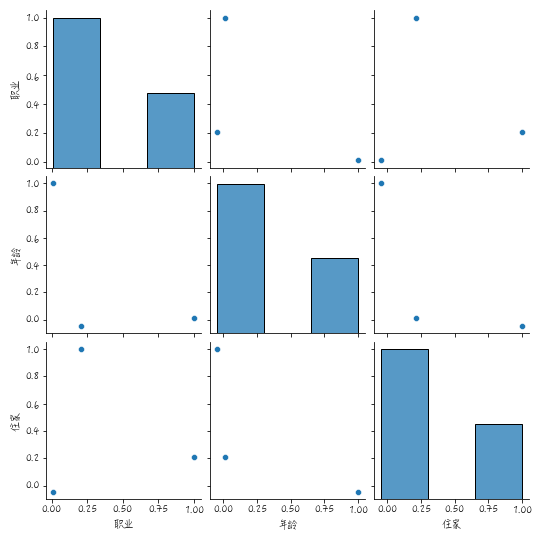




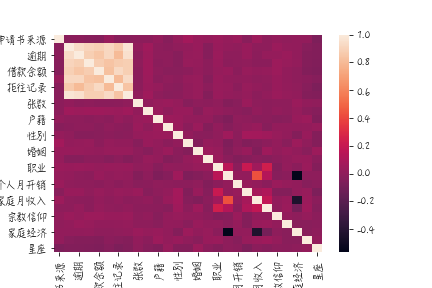
1. 由常识得，收入风险情况特征应与职业、年龄、住家有关，验证得：







1. 可视化呈现结果（给出可视化呈现界面图，并分析结果做说明）

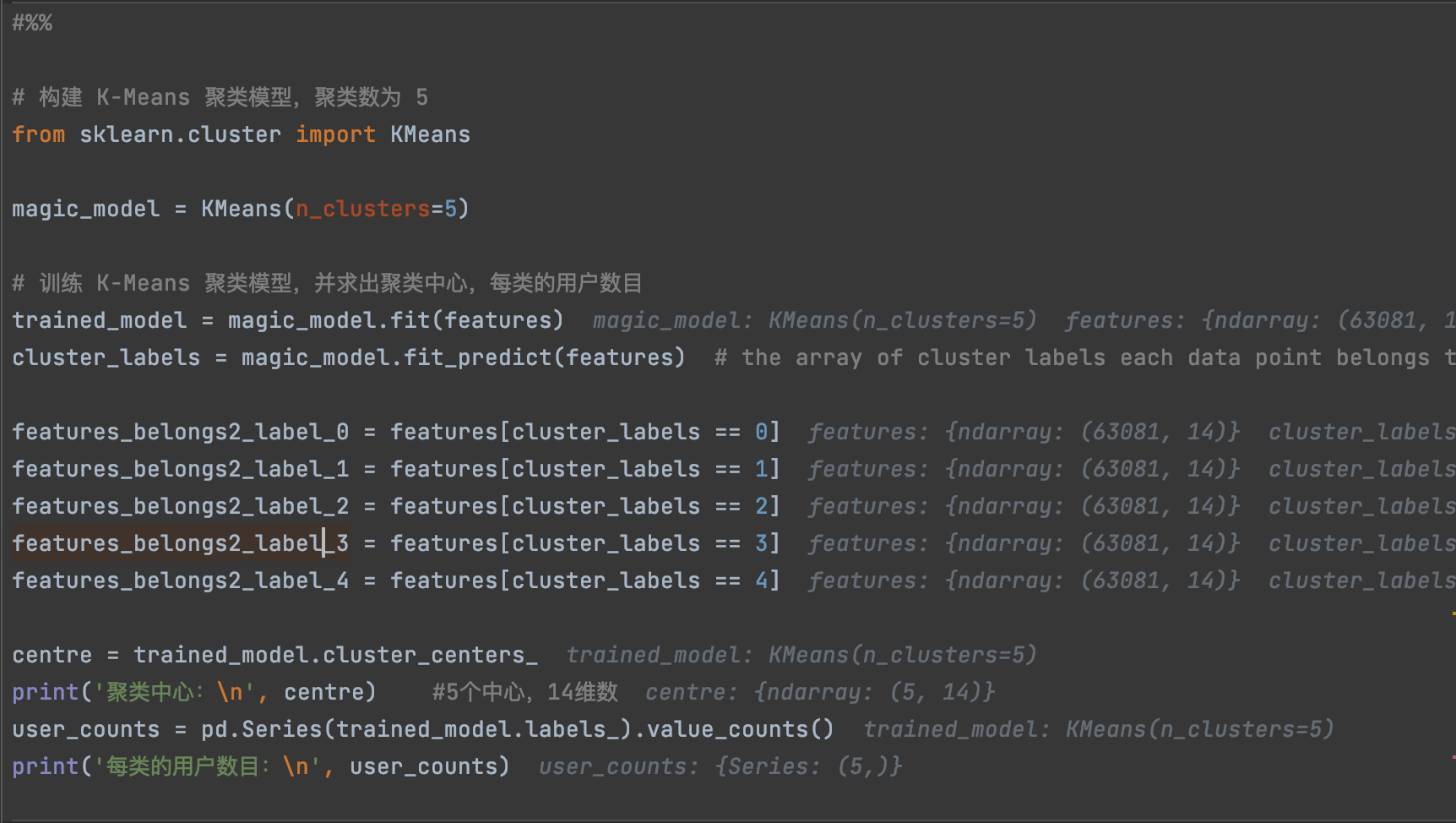


可以看到有两个发亮的区域，发亮代表相关性比较强，可以从这个图中得到启发，结合常识选择用来构建融合特征的原始特征。

**四、数据分析模型**

1. 结合分析的目标，拟采用哪一种模型（如聚类、分类、回归）开展分析

采用模型：聚类模型

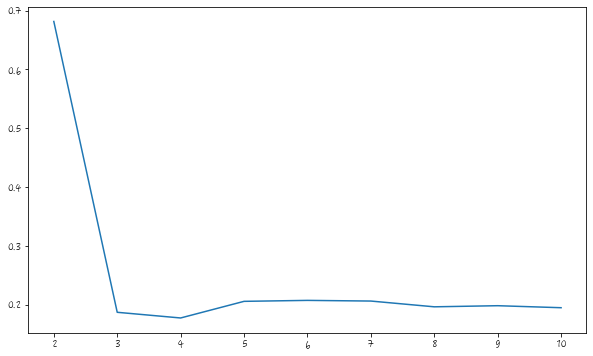


理由：这是按照需求来的，我最终的目的是构建信用卡高风险客户识别的模型，那么就需要根据我构建的3个特征对客户进行分群，所以采用聚类算法。

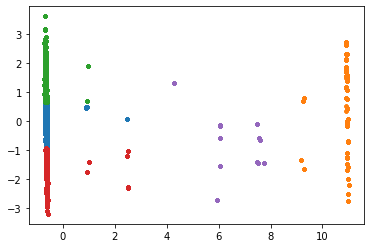
1. 模型评估

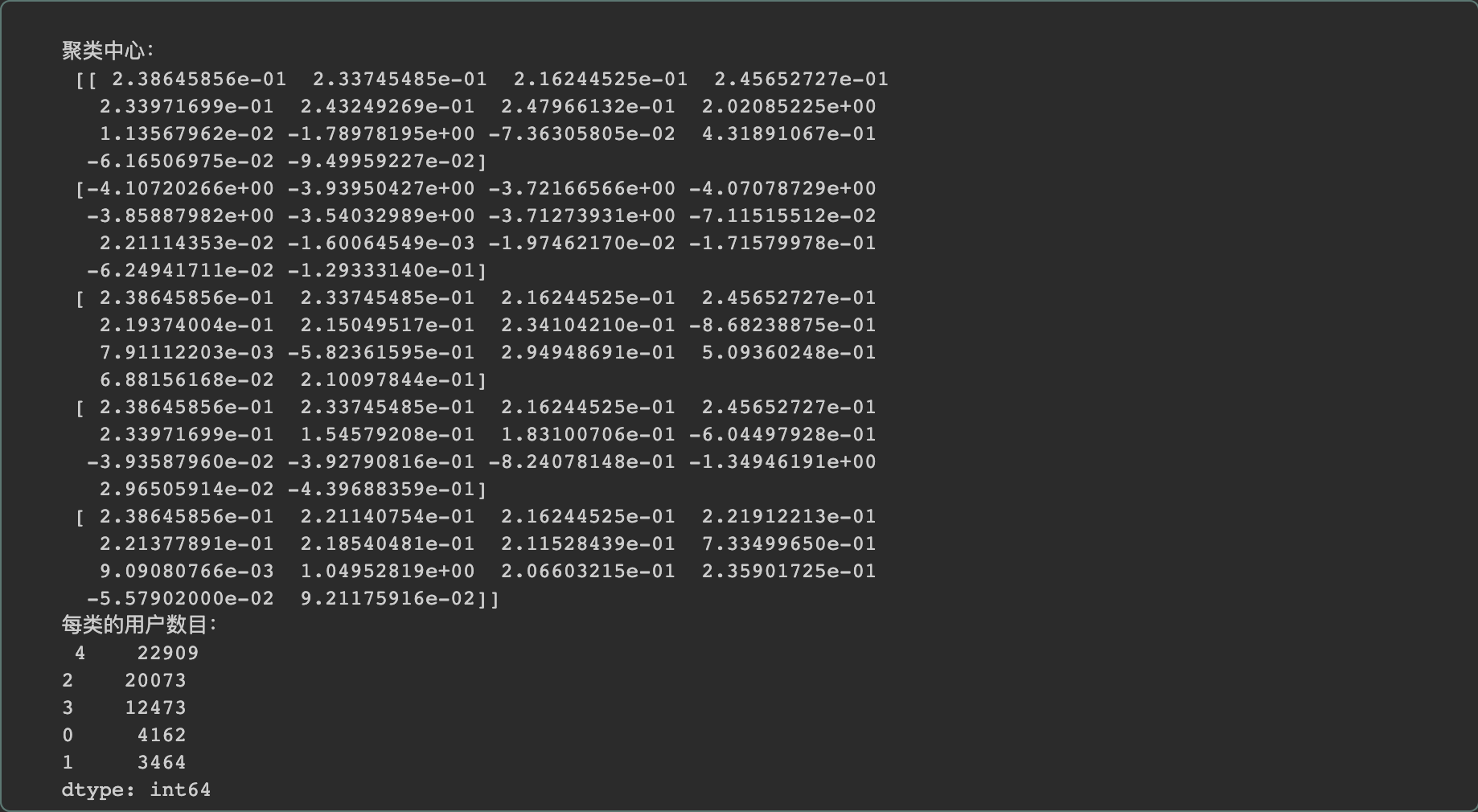
评估方法：轮廓系数





1. 可视化结果





分为五类，由图可见类别划分还是比较清晰的。

**五、方案评估**

达到了预期目标，完成了设计任务，难度中等。