|  |  |
| --- | --- |
| **字段名称** | **描述** |
| **MONTH\_ID** | 数据所属的月份标识（格式：YYYYMM） |
| **USER\_ID** | 用户唯一标识符 |
| **INNET\_MONTH** | 用户在网月数，即用户从开通服务到当前月份的累计月份数 |
| **IS\_AGREE** | 是否同意某项服务或协议（1表示同意，0表示不同意） |
| **AGREE\_EXP\_DATE** | 同意协议的到期日期（格式：YYYYMM） |
| **CREDIT\_LEVEL** | 用户信用等级，通常用数值表示信用评分或等级 |
| **VIP\_LVL** | 用户VIP等级，表示用户的VIP级别 |
| **ACCT\_FEE** | 用户账户月费，单位为货币（如元） |
| **CALL\_DURA** | 通话总时长，单位为秒 |
| **NO\_ROAM\_LOCAL\_CALL\_DURA** | 无漫游本地通话时长，单位为秒 |
| **NO\_ROAM\_GN\_LONG\_CALL\_DURA** | 无漫游国内长途通话时长，单位为秒 |
| **GN\_ROAM\_CALL\_DURA** | 国内漫游通话时长，单位为秒 |
| **CDR\_NUM** | 通话记录数量 |
| **NO\_ROAM\_CDR\_NUM** | 无漫游通话记录数量 |
| **NO\_ROAM\_LOCAL\_CDR\_NUM** | 无漫游本地通话记录数量 |
| **NO\_ROAM\_GN\_LONG\_CDR\_NUM** | 无漫游国内长途通话记录数量 |
| **GN\_ROAM\_CDR\_NUM** | 国内漫游通话记录数量 |
| **P2P\_SMS\_CNT\_UP** | 用户发送的点对点短信数量 |
| **TOTAL\_FLUX** | 总流量使用量，单位为GB |
| **LOCAL\_FLUX** | 本地流量使用量，单位为GB |
| **GN\_ROAM\_FLUX** | 国内漫游流量使用量，单位为GB |
| **CALL\_DAYS** | 通话天数，即用户在一个月内进行通话的不同天数 |
| **CALLING\_DAYS** | 主叫天数，即用户在一个月内作为主叫发起通话的不同天数 |
| **CALLED\_DAYS** | 被叫天数，即用户在一个月内作为被叫接收通话的不同天数 |
| **CALL\_RING** | 通话铃声数，可能指通话过程中铃声的总次数或时长（需根据具体定义确认） |
| **CALLING\_RING** | 主叫铃声音数，可能指主叫过程中铃声的总次数或时长（需根据具体定义确认） |
| **CALLED\_RING** | 被叫铃声音数，可能指被叫过程中铃声的总次数或时长（需根据具体定义确认） |
| **CUST\_SEX** | 客户性别（1表示男性，2表示女性，具体编码需参考数据字典） |
| **CERT\_AGE** | 客户认证年龄，表示客户的年龄 |
| **CONSTELLATION\_DESC** | 星座描述，表示客户的星座 |
| **MANU\_NAME** | 设备制造商名称 |
| **MODEL\_NAME** | 设备型号名称 |
| **OS\_DESC** | 操作系统描述，表示设备的操作系统版本 |
| **TERM\_TYPE** | 终端类型，通常用数值或代码表示不同的终端类别（需参考数据字典） |
| **IS\_LOST** | 是否流失标识（1表示流失，0表示未流失） |

（1）对原始数据进行预处理，其中包括：查找并删除重复数据；删除手机品牌、手机型号和操作系统描述这3个属性，降低数据集的维度；将每个客户信息处理为一行数据；处理缺失值与异常值

**处理缺失值**

删除所有包含缺失值的行

**处理异常值**

对数值型列进行异常值检测和处理。

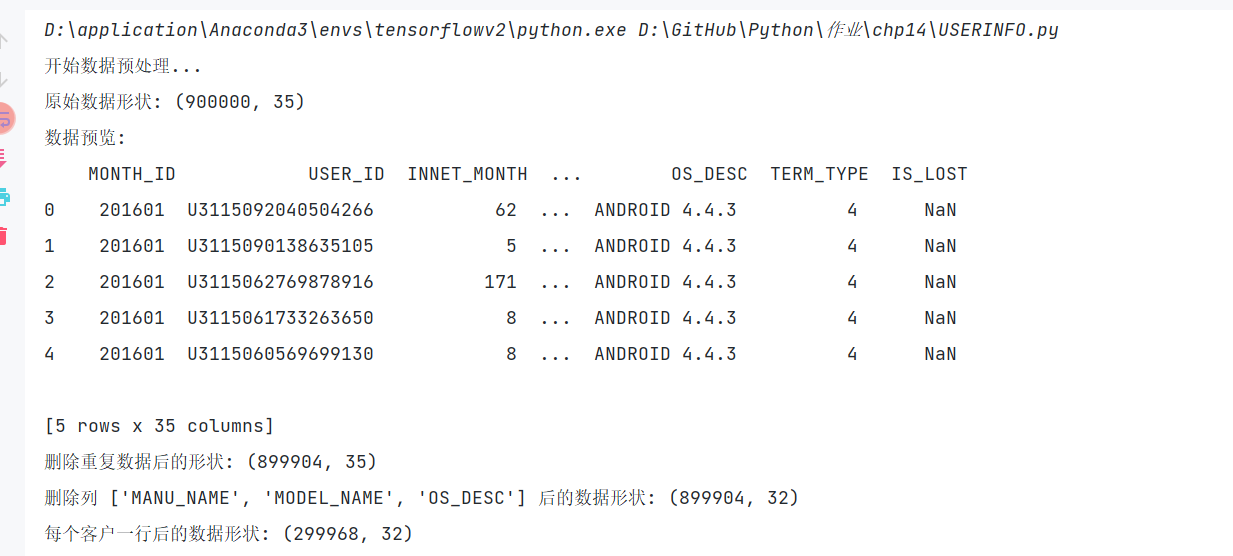
根据列的偏度（skewness）选择不同的方法：

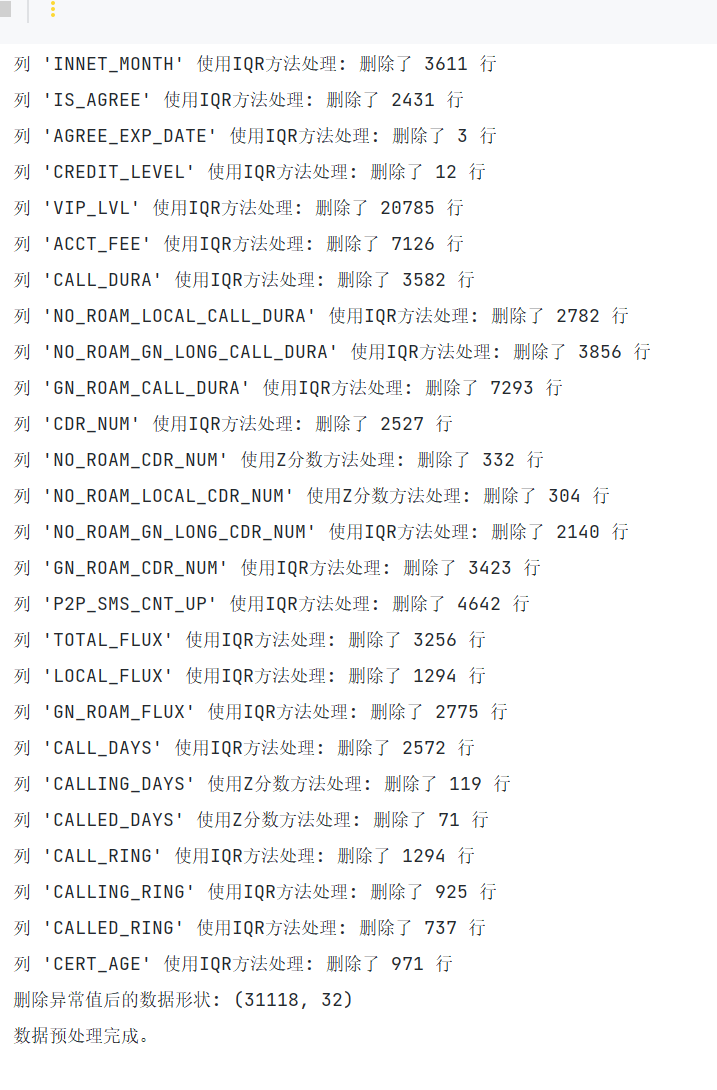
① 偏度较小（abs(skew) < 1），使用Z分数(Z-score)方法：

将Z分数绝对值大于3的数据视为异常值并删除。

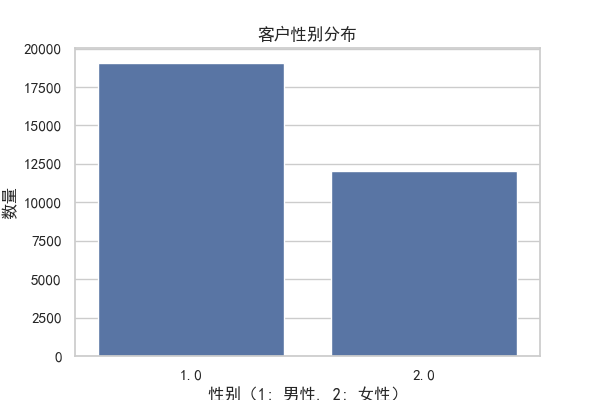
② 偏度较大，使用四分位距法(IQR)：

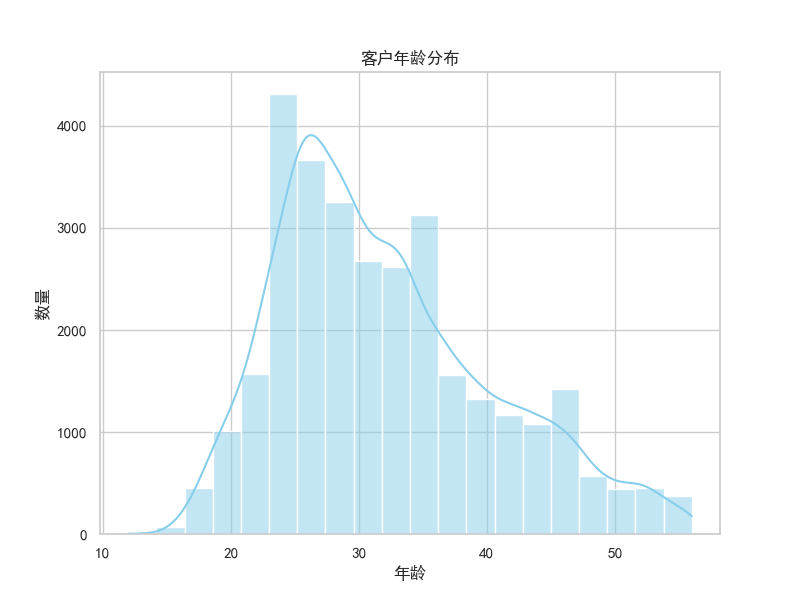
计算IQR = Q3 - Q1，并将小于Q1 - 1.5IQR或大于Q3 + 1.5IQR的数据视为异常值并删除。

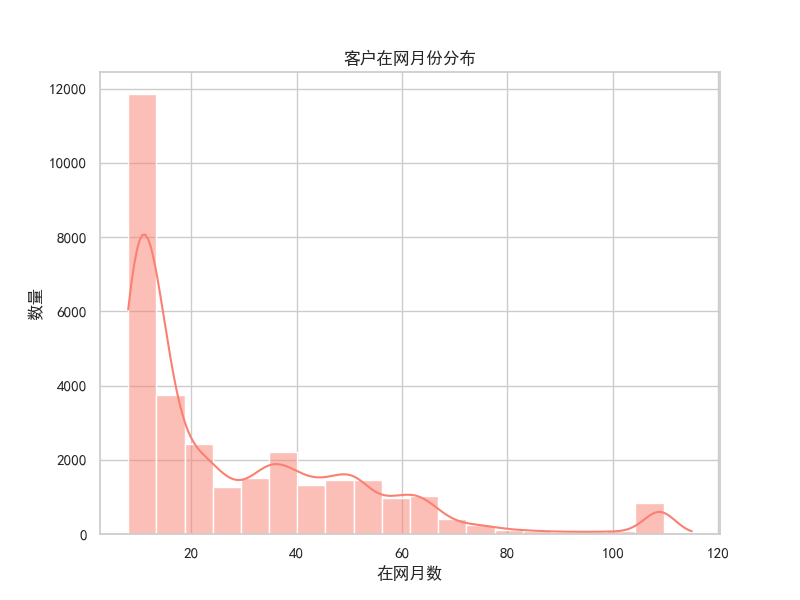


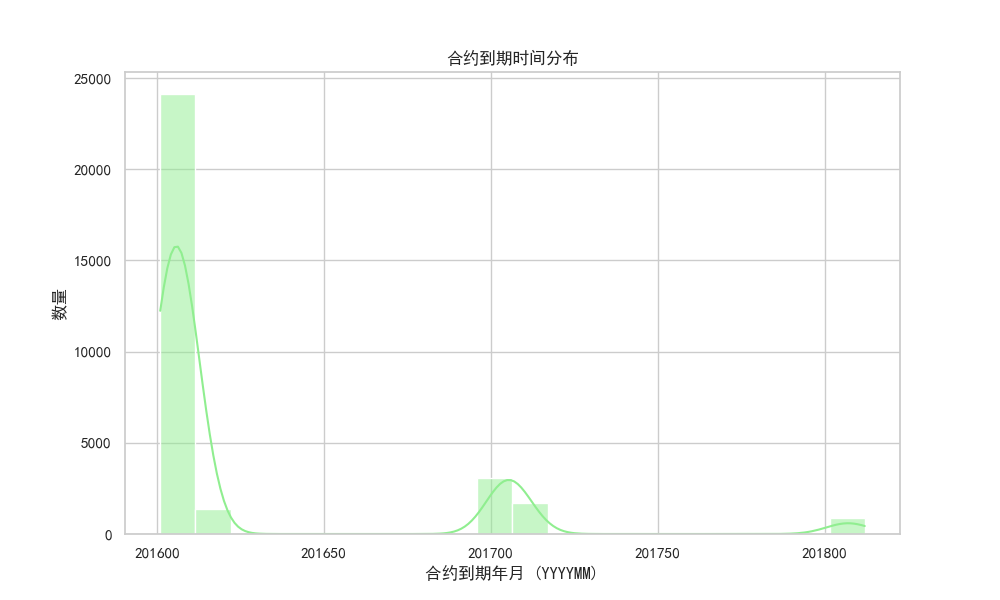


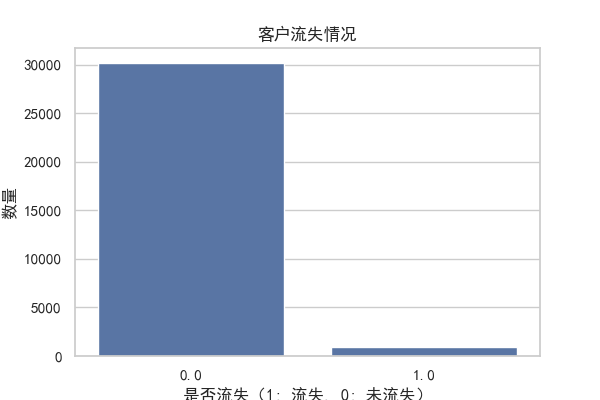
（2）对客户的性别、年龄、在网市场、合约到期时间、客户是否有效、信用等级进行分析

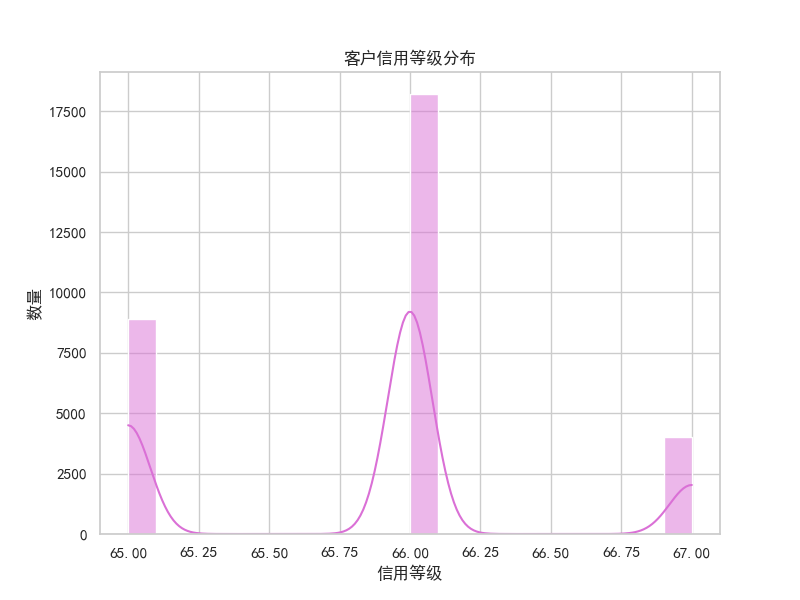




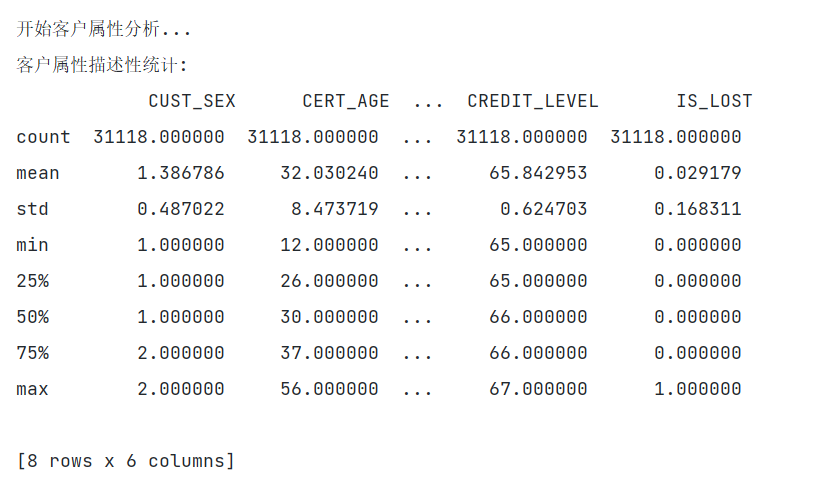




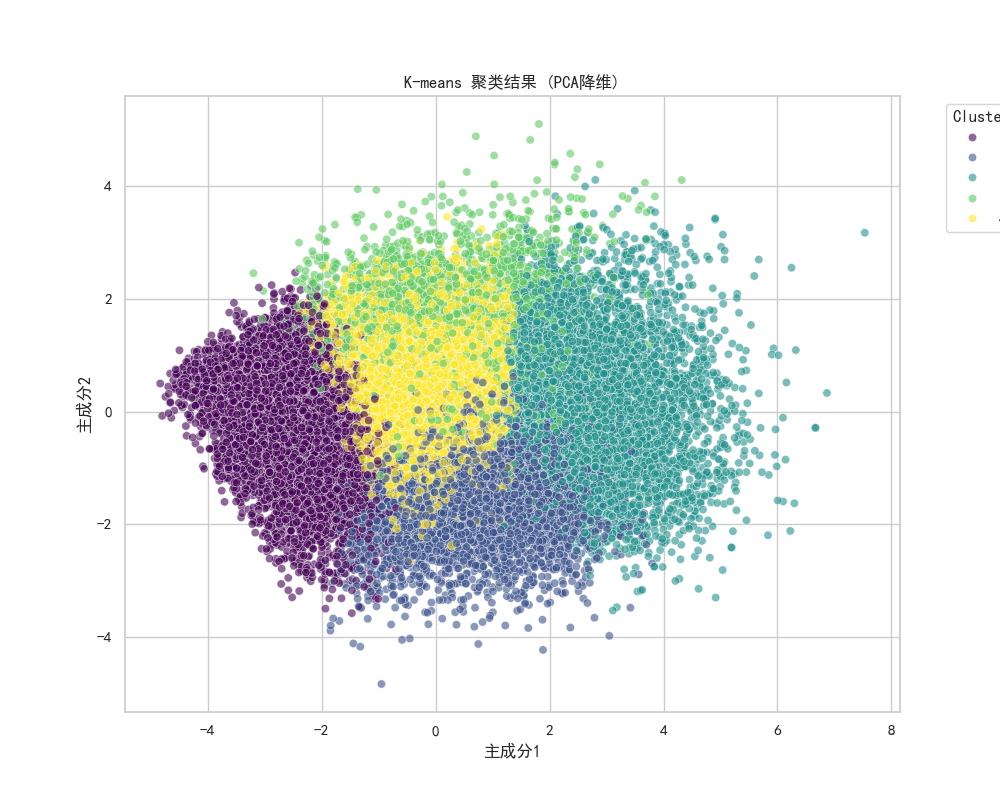




使用 data.describe() 对关键的数值属性进行基本统计描述（如均值、中位数、标准差、最大值、最小值等），帮助我们更好地了解数据的分布和特征。



（3）聚类



目前已经将客户分为5类并得到各簇特征中心，但聚类质量一般。这为后续工作提供了改进方向：需要进一步优化特征选择、预处理方式或重新选择K值，以期获得更清晰、更有业务解释力的客户群划分。