地学大数据课程项目

软件学院1953871 邓泉

指导教师：李江峰

1. **技术栈**

****

编程语言：python

开发工具：jupyter notebook

数据处理：numpy，pandas

机器学习：scikit-learn

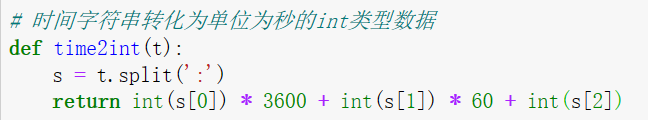
数据可视化：matplotlib

1. **用户基本行为的统计**
2. **数据清洗**

原始数据存在一些问题：

1. end\_time 通话结束时间存在错误数据0:00:00
2. raw\_dur 通话时长存在错误数据，与 (end\_time – start\_time) 不一致

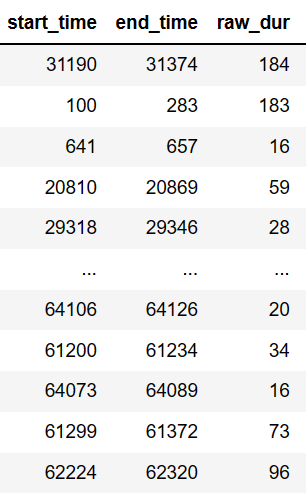
对应的解决方案：



利用time2int 将 start\_time end\_time 从字符串类型转化为int类型数据，方便后续计算。

1. 用(start\_time + raw\_dur) % day\_time（***day\_time = 24 \* 3600 单位：s***）代替 end\_time 通话结束时间
2. 计算 (end\_time - start\_time + day\_time) % day\_time作为raw\_dur通话时长

处理结果局部展示：

****

1. **<主叫号码, 每日平均通话次数, 每日平均通话时长、每个通话平均时长>**

处理步骤如下：

1. 使用 Groupby 根据主叫号码进行分组



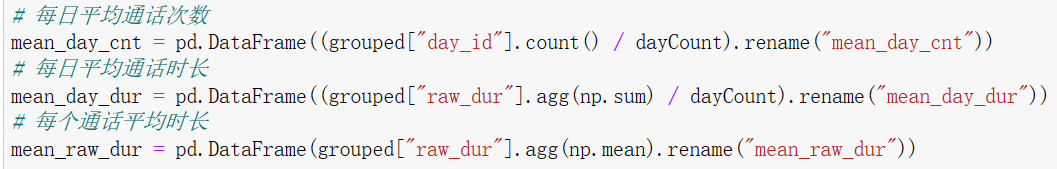
1. 计算出总天数：29



1. 分组聚合函数计算

计算原理公式：

* **每日平均通话次数 = 总通话次数 / 总天数**
* **每日平均通话时长 = 总通话时长 / 总天数**
* **每个通话平均时长 = 总通话时长 / 总通话次数**

****

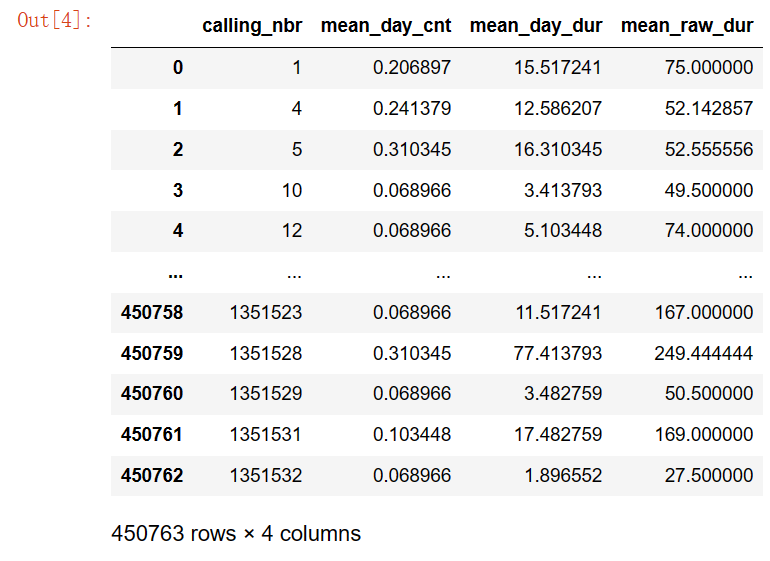
1. 融合数据



1. 保存



结果局部展示：



1. **<主叫号码, 时间段1占比, ..., 时间段8占比>**

代码：



处理步骤如下：

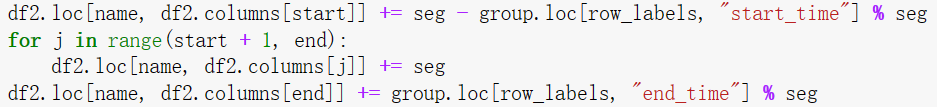
1. 创建DataFrame，初始化columns，其中all\_dur为总通话时长，t1 - t8 为每个时间段的

****

1. 根据主叫号码进行分组，可以得到每个主叫号码下对应的所有通话记录，累积计算每一条通话记录到对应时间段的时长

每条通话记录的计算方法如下：

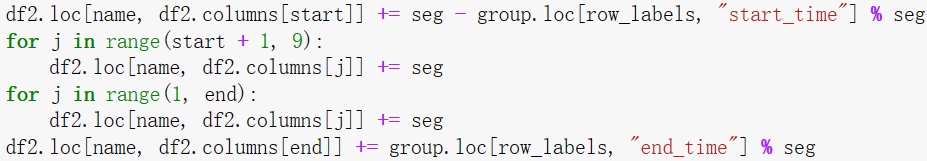
1. 计算出start\_time和end\_time所处时间段（1 - 8）start和end
2. 根据start和end的大小关系分类讨论：
   1. start < end：



* 1. start == end：



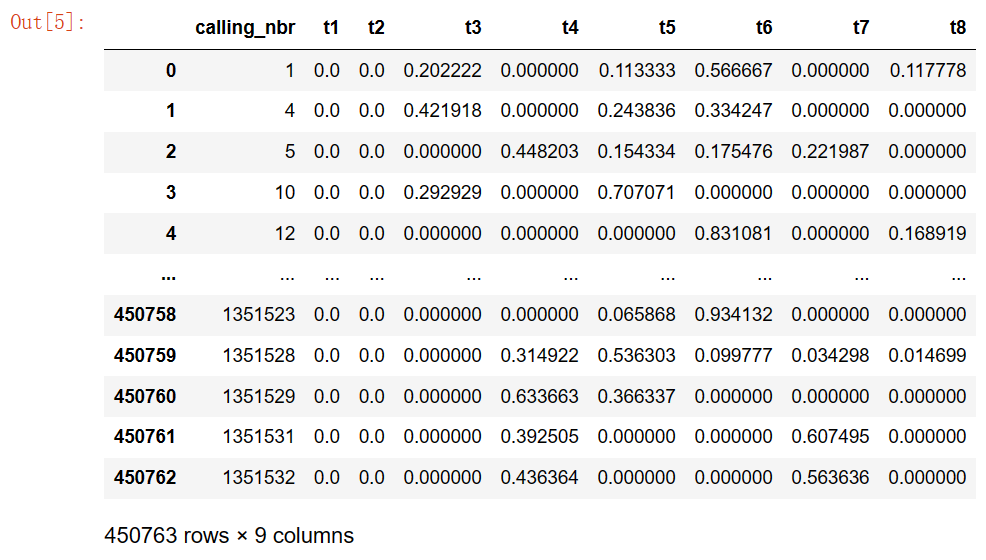
* 1. start > end：通话跨过了次天零点



1. 每个时间段的秒数除以总秒数，最终得到用户在各个时间段通话时长所占比例

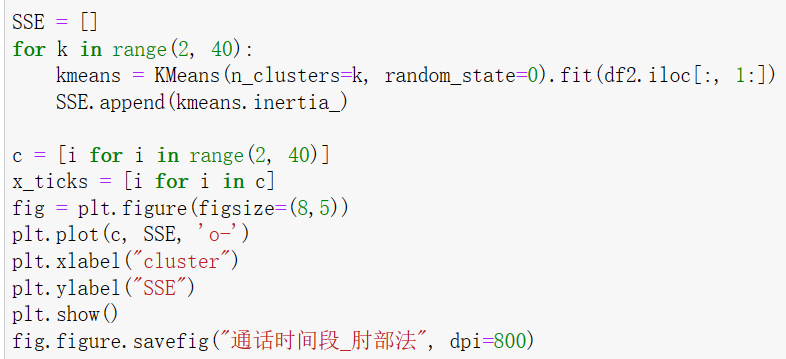


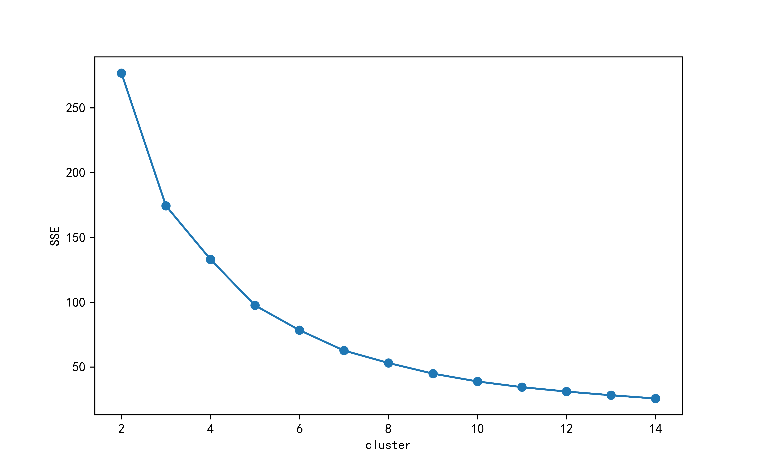
结果局部展示：



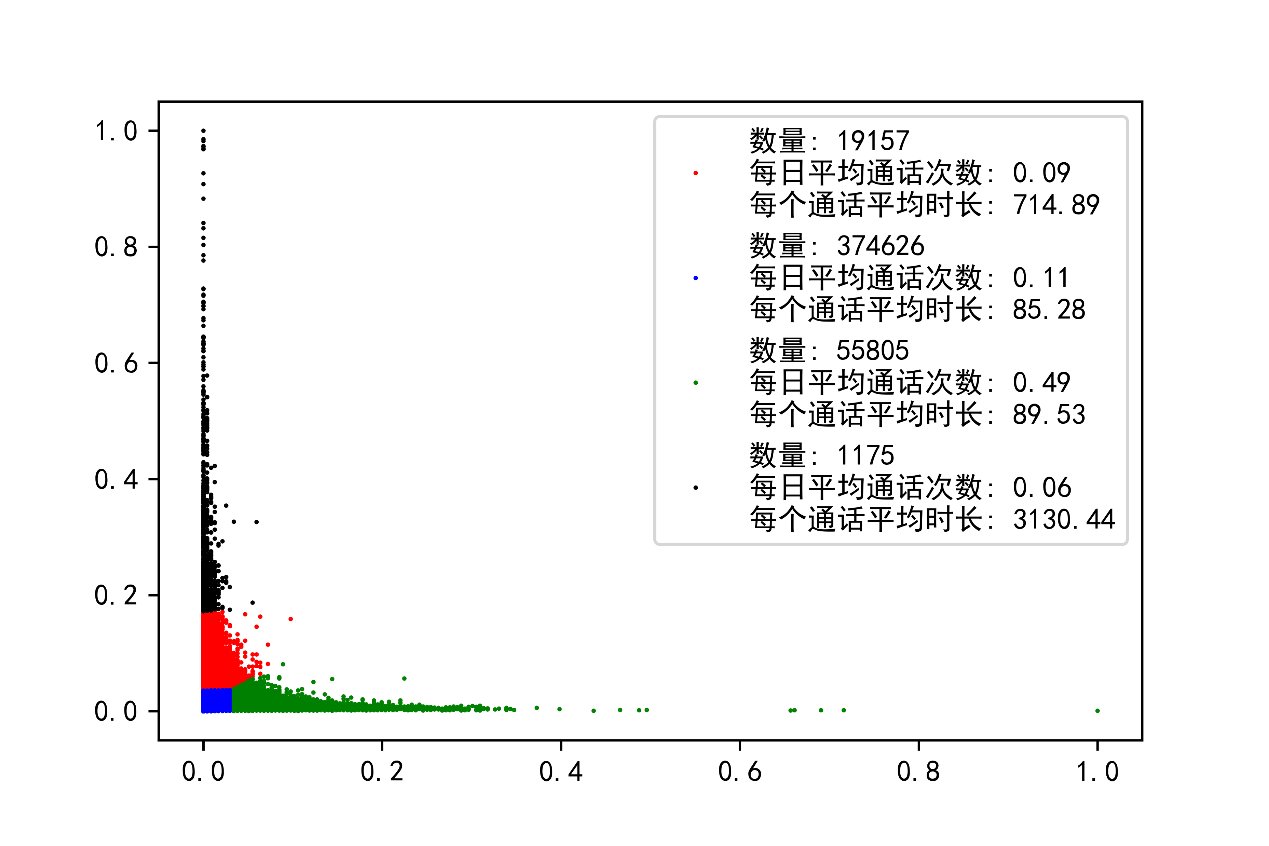
1. **用户行为特征深度分析**
2. **<每日平均通话次数， 每个通话平均时长>通话频率和时长**

对用户的***<每日平均通话次数， 每个通话平均时长>***进行K-Means聚类分析，并利用肘部法确定最佳的cluster数量为4。





聚类结果分析：分别以***<每日平均通话次数， 每个通话平均时长>***为横纵坐标绘图

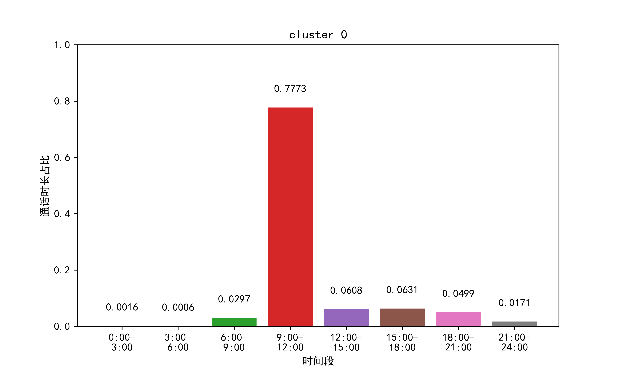
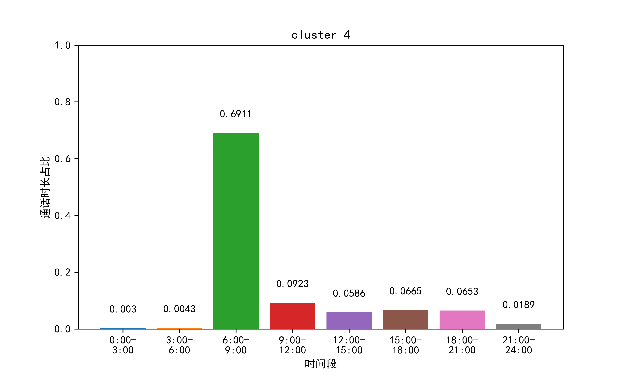
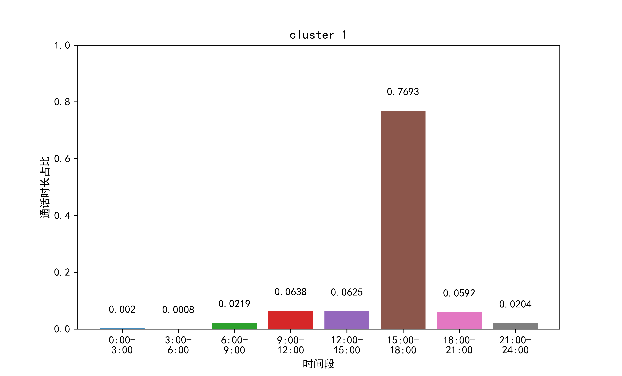
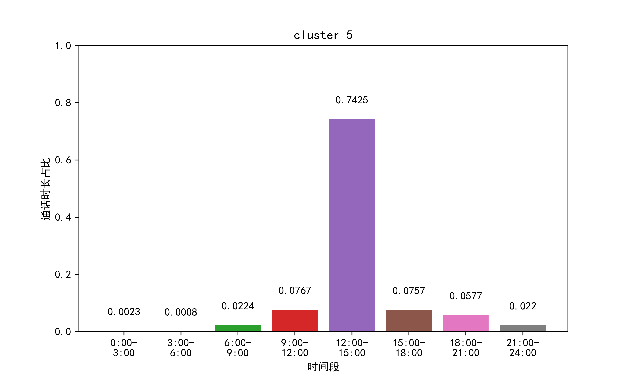
****

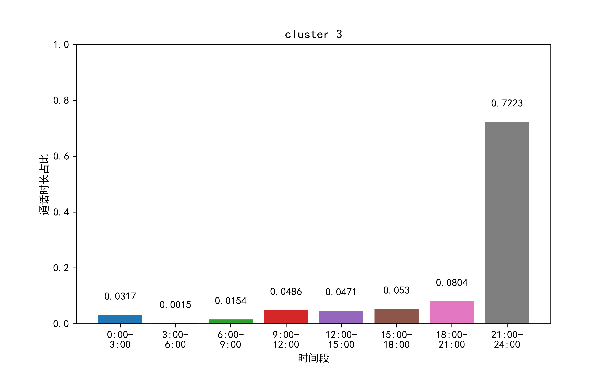
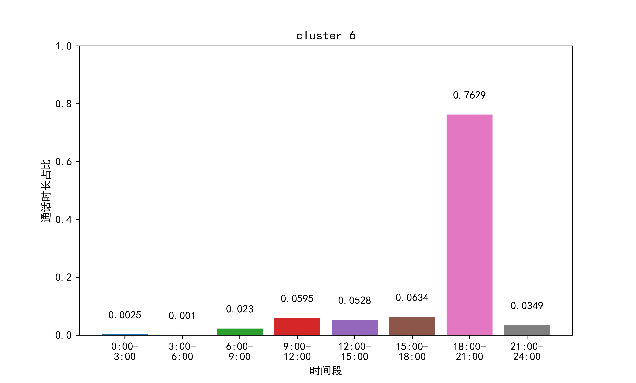
* 黑色类：人数最少，通话次数少，通话时间长
* 蓝色类：人数最多，通话次数少，通话时间短
* 红色类：相比于黑色类人群通话次数稍稍增多，通话时长介于黑色类与蓝色类之间
* 绿色类：通话次数多，通话时间短

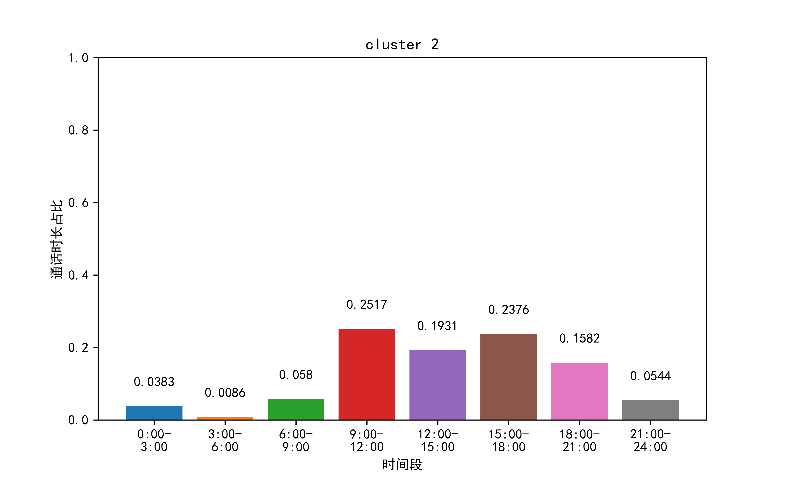
1. **通话时刻（通话时间段）**

对用户的八维特征向量***<时间段1占比, ..., 时间段8占比>***进行K-Means聚类分析，根据肘部法确定cluster数量为7。

聚类结果分析：对于每一个簇，将该簇下对应的所有主叫号码的通话时间段和通话时长做统计分析，绘制直方图表示每个簇的通话时长分布情况。





结果显示，其中1个簇反映的特征是该类人群通话时长占比在各个时间段相对均匀；其余6个簇的特征是有一个时间段的通话时长占比特别突出，说明该类人群在对应的时间段内经常通话，而除了这6个簇对应的“最常通话时间段”之外剩下的2个时间段是0:00-3:00,3:00-6:00，这恰好是人们的睡眠时间段，“最常通话时间段”一般不会产生在这两者里面，这样的数据分析结论也非常符合实际。