**实验五 pandas数据预处理**

**【实验目的】**

1. 掌握使用pandas库对重复数据、缺省数据和异常数据进行处理的方法。
2. 掌握使用pandas库非数值数据的数值化处理。
3. 掌握使用pandas库进行数据标准化的操作。

**【实验内容】**

1. pandas常规操作；
2. 使用pandas进行数据预处理；

**【预备知识】**

1. pandas库
2. Python编程基础

**【实验设备】**

1.硬件：PC机

2.软件：python3开发环境

**【实验步骤】**

1. 使用pandas读取‘实验5数据.xlsx’文件的数据，并保存在data中；
2. 将data中有缺省值的行删除掉；
3. 将获赞数、播放数、平均播放数、平均获赞数、总粉丝人数、充电人数和作品数中小于等于0所在的行删除掉；
4. 将性别列中的三种不同取值分别用数值替换，其中男替换成0、女替换成1、保密替换成2；
5. 分析data中各列数据的相关性，如果有两列的相关性大于0.7，则将这两列打印出来。
6. 使用箱线图的方法，将粉丝数列的异常值打印出来。
7. 使用离差标准化对获赞数列进行数据标准化。
8. 使用标准差标准化对播放数列进行数据标准化。
9. 定义：获赞效率为平均每1000次播放的获赞数，为data添加获赞效率列；
10. 定义：充电效率为平均每1000次播放的充电数，为data添加充电效率列；
11. 定义：作品效率为平均每个作品的充电数和点赞数的乘积，为data添加作品效率列；
12. 自定义一种方法，评选出不同分组的最有价值UP主。