**北京邮电大学软件学院**

**2018－2019学年第二学期实验报告**

**课程名称： 云计算数据中心**

**项目名称： 数据分析**

**项目完成人：**

**姓名：\_\_\_白烨淞\_\_\_\_\_学号：\_\_2016522050\_\_\_\_\_\_**

**指导教师： 管皓**

**日 期： 2018 年 12 月 07 日**

1. **实验目的**

学生通过使用编程语言或数据分析软件巩固大数据分析方法

1. **实验内容**

**1 常用统计量**

**（1）随机生成1000个 [0,10000] 范围内的整数。**

**（2）求出数据中的中均值，众数，中位数，极差，标准差，变异系数。**

**（3）对这些数据进行预处理，使其均值为零，方差为1。**

**（4）对（3）中预处理后的数据进行可视化显示（曲线形式，直方图形式等）**

**2 神秘的女孩**

**现有两幅相似图像的数据（矩阵形式存储于1.xlsx, 2.xlsx），之所以相似，是因为记录了同一场景，差别在于其中一个场景中出现了一位神秘的女孩。**

**(1) 请仅根据两份数据，经过分析，判断神秘的女孩在哪一个图像内。请说明自己判断的依据，及分析过程。**

**(2) 根据两份数据，对图像进行可视化，验证自己的判断。**

**(3) 请从你的直觉上为两幅图像的相似性打分（0~1之间，完全不像为0，完全一样为1），计算两幅图像的相关系数，你觉得计算出来的结果与你的主观评价是否相近。**

1. **实验环境**

编译环境：Pyhton3

操作系统：Windows10

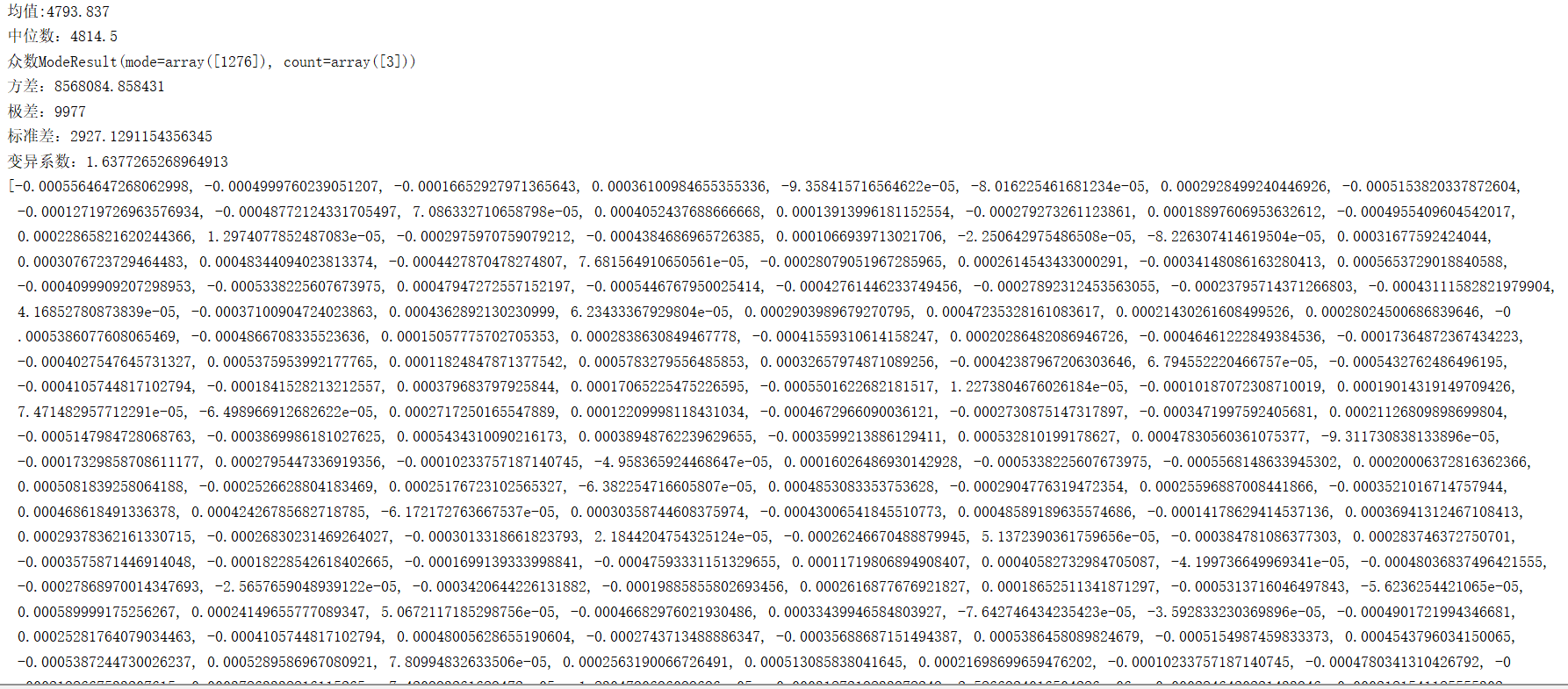
编译器：Pycharm2018.1.1

1. **实验结果**
2. **常用统计量**

（随机生成的1000个整数（部分））



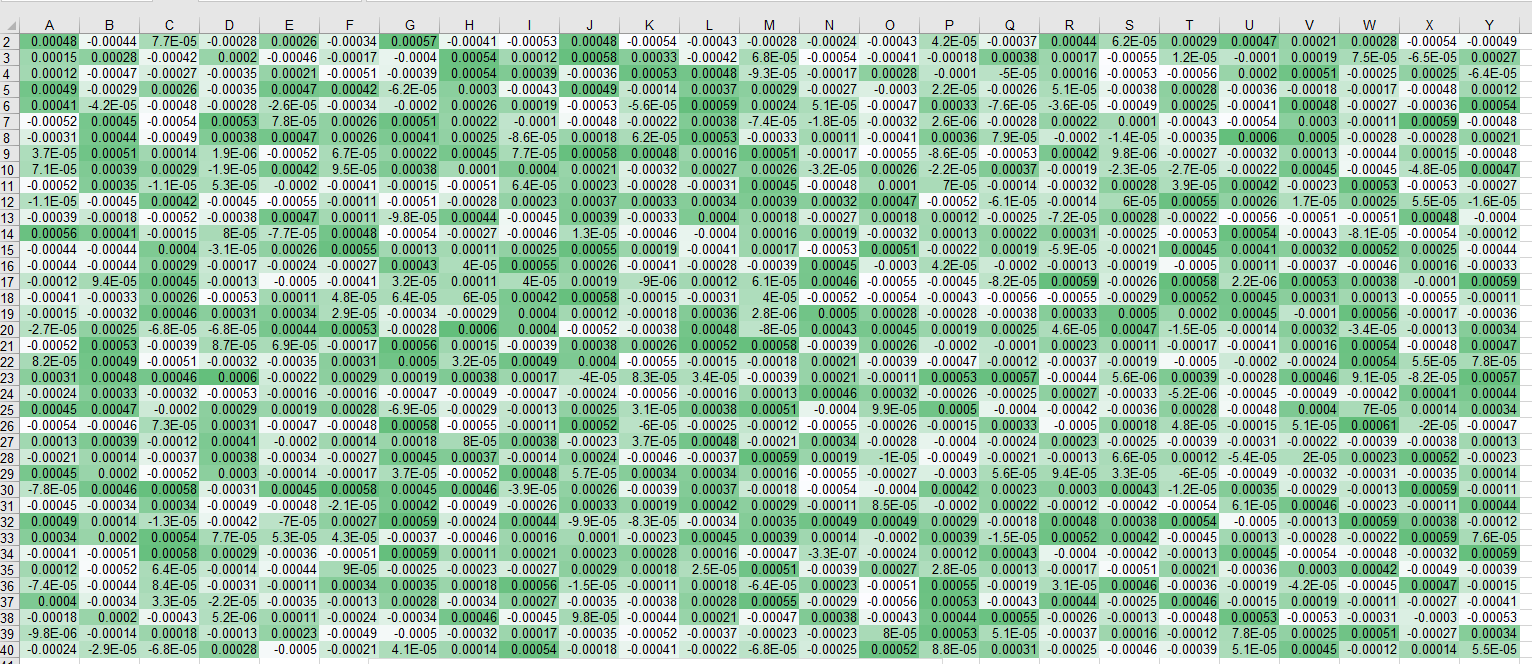
（经零-均值转化过后的数据及相关参数（部分））



（可视化结果）

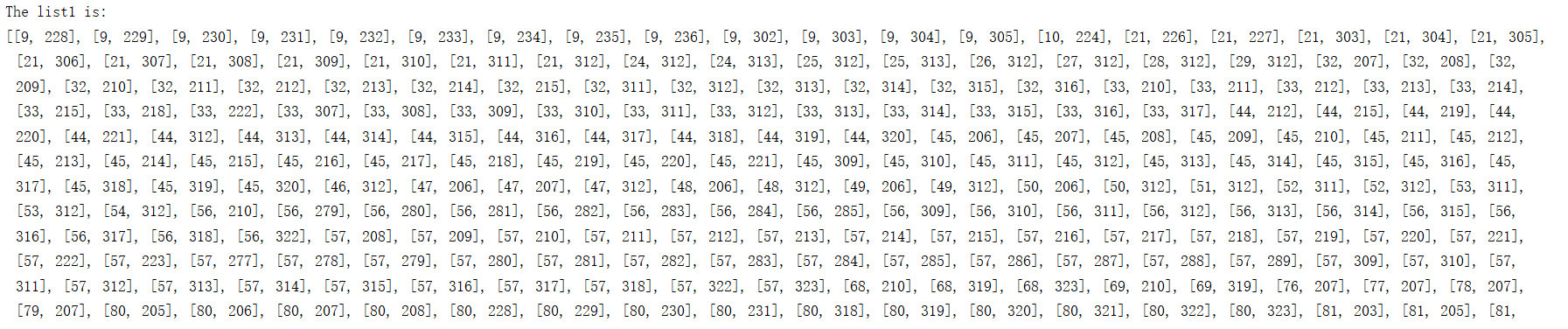
统计表形式

色阶表示法

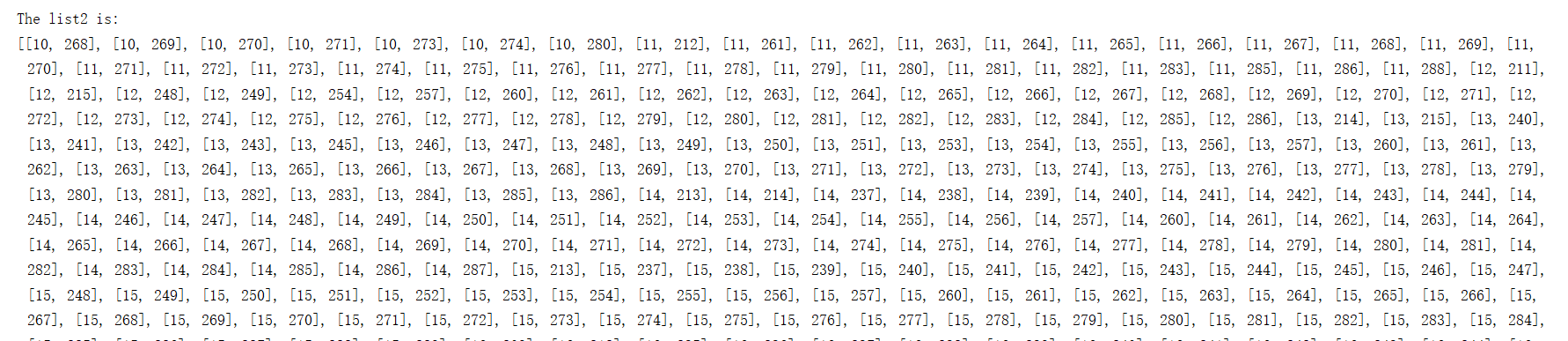


1. **神秘的女孩**

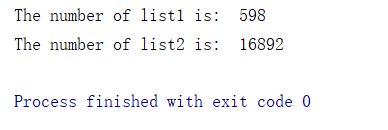
（1.xlsx中元素多于相应2.xlsx中元素值100的坐标（部分））



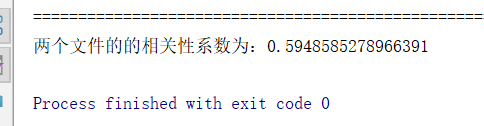
（2.xlsx中元素多于相应1.xlsx中元素值100的坐标（部分））



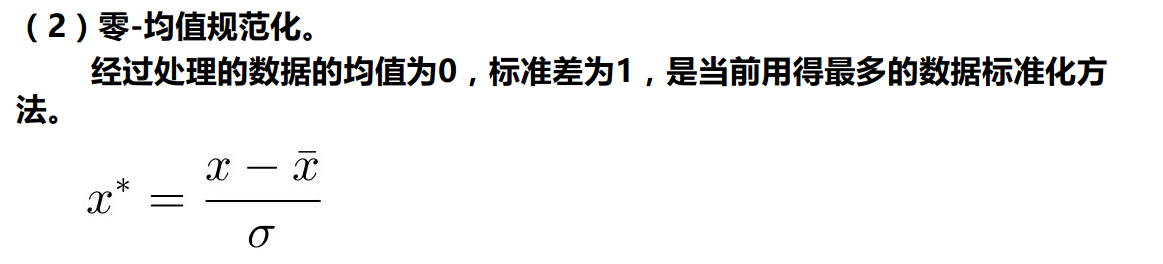
（两个列表的元素个数）



（两个文件的相关性系数）

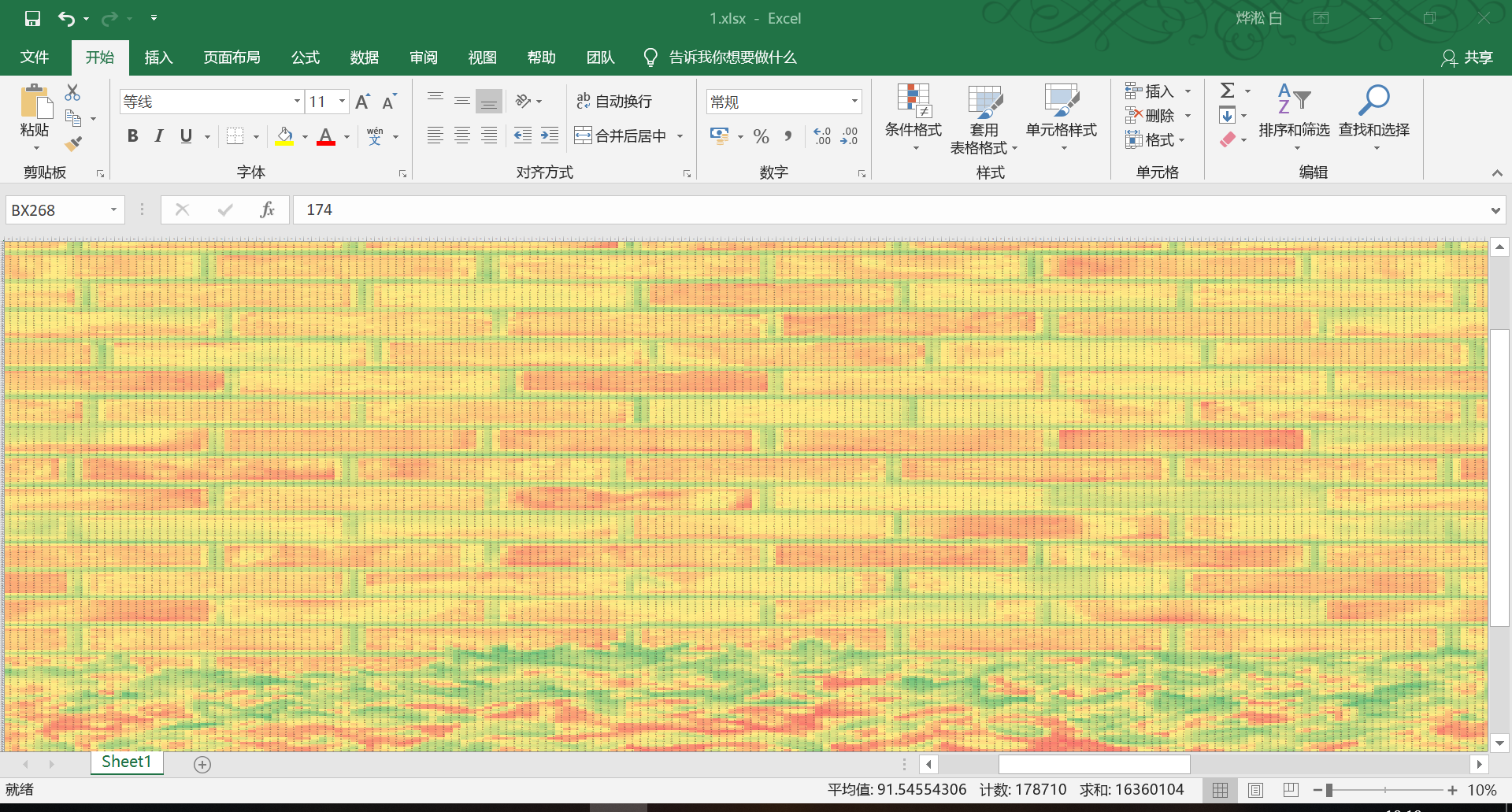


1. **附录**
2. **常用统计量**
3. 采用python的random库进行随机数的生成，将生成的1000个数存入一个列表（详见结果附件）
4. 通过python扩展库numpy和scipy进行数据的统计量计算，两个扩展库提供了很多统计量函数（具体函数详见代码文件）
5. 进行零-均值规范化，通过循环遍历每一个列表的数据

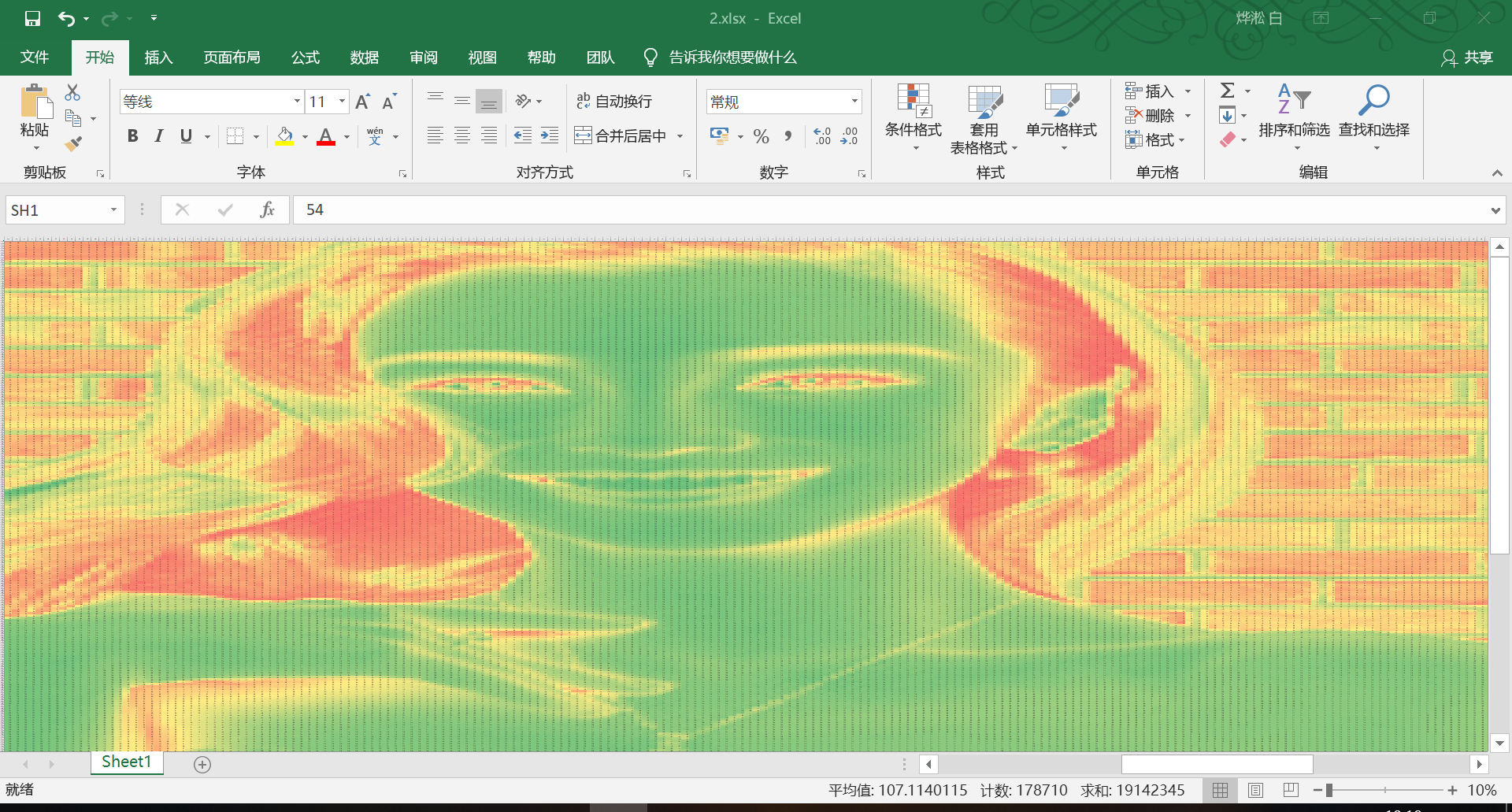


1. 为了进行数据可视化，首先使用扩展库xlwt将列表数据导出至xls表格（After\_data.xls详见附件），在Excel工具中，我们采用堆积柱形图以及单元格色阶变化显示的形式进行数据可视化
2. **神秘女孩**
3. 我们首先推测这一组二维单元数值型数据中，每个数值代表一个颜色，不同的颜色值（像素块）组成了画面，我们再退一步思考，不如将表格中的数据值看作颜色深浅度形成单色调的色阶图，我们通过Excel表格来实现这一功能，而在代码上我们对每个相对应的的单元格进行比较，分别取得大于另一个相对应文件100的单元格坐标分别存储在list1和list2中，我们发现list2中的数据量远大于list1中的数据量，据此我们推测神秘的女孩在2.xlsx中
4. 我们分别对两个表格进行色阶表示（开始-条件格式），结果如下

（1.xlsx）

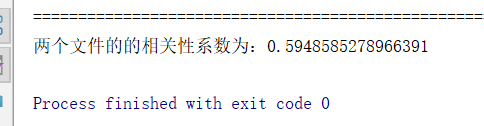


（2.xlsx）



神秘的女孩，一目了然。

（4）在我的主观上我给两幅图片的相似度0.65分，因为女孩图形化的占比大概在50%左右。再python中我们使用pandas扩展库来计算相关性系数（用法详见代码）



我觉得确实是比较相近的