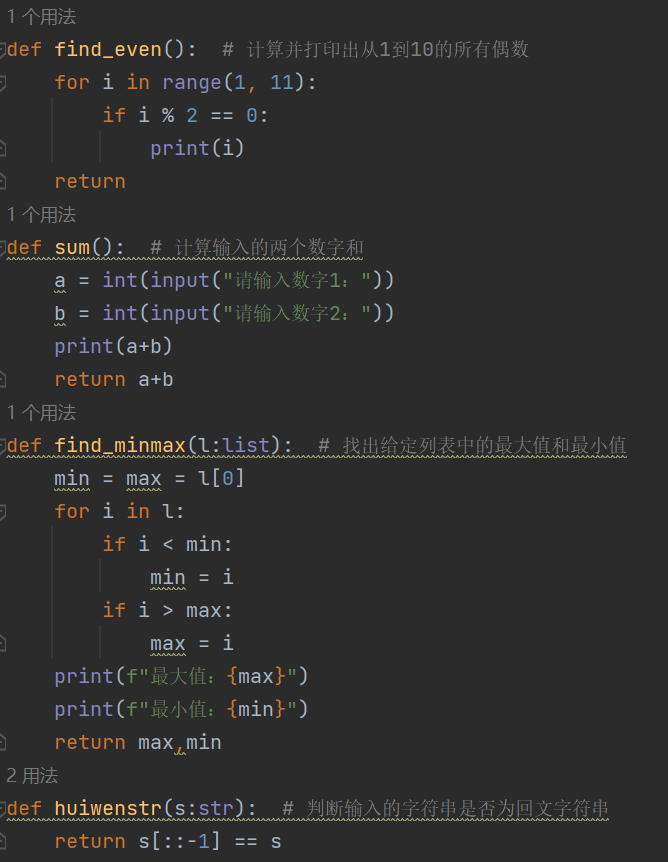
生物信息Perl编程上机实验报告

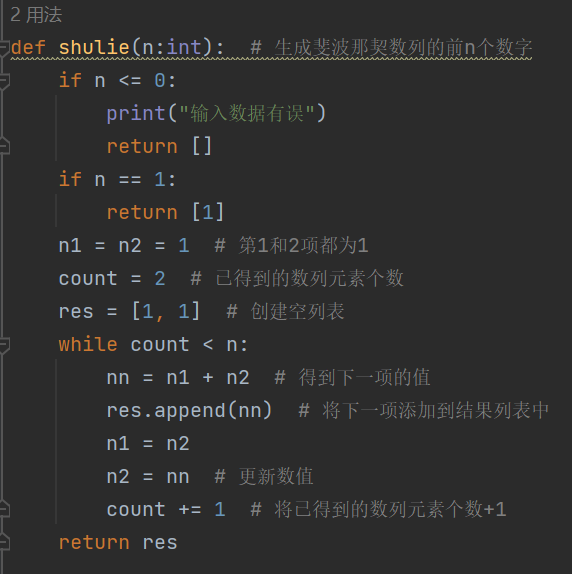
生信基地2201班 王天皓 U202212960

# 第一次上机

## 1. 基础题目

**实现代码：**

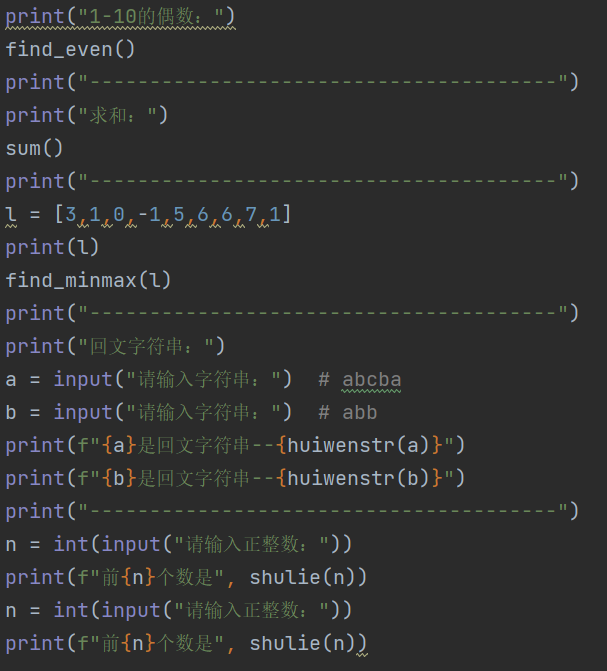




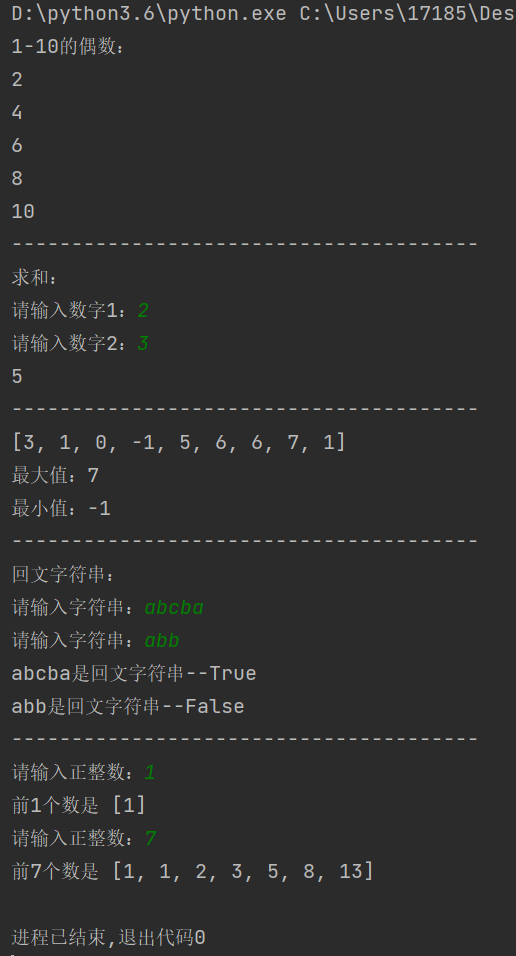
（报告中所有实现代码的文字版都在

<https://github.com/lwstkhyl/python/tree/main/%E7%94%9F%E4%BF%A1%E7%BC%96%E7%A8%8B%E4%B8%8A%E6%9C%BA%E4%BD%9C%E4%B8%9A>中）

**测试代码：**

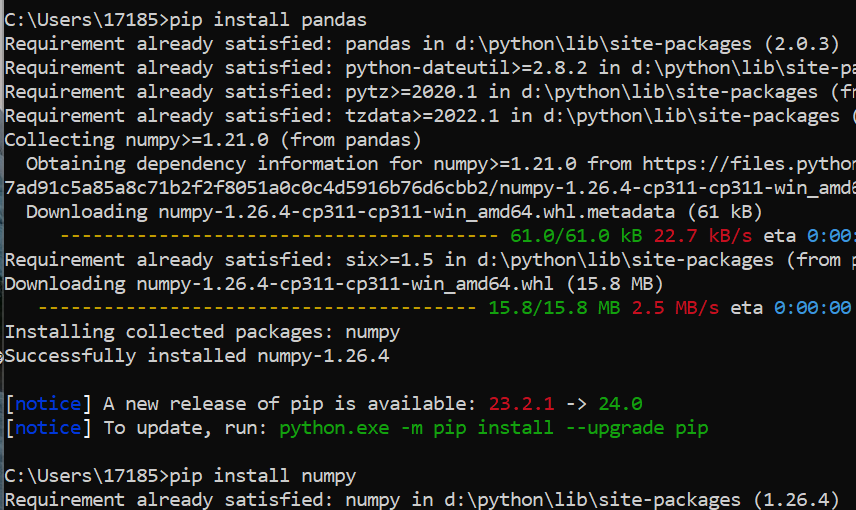


**测试结果：**

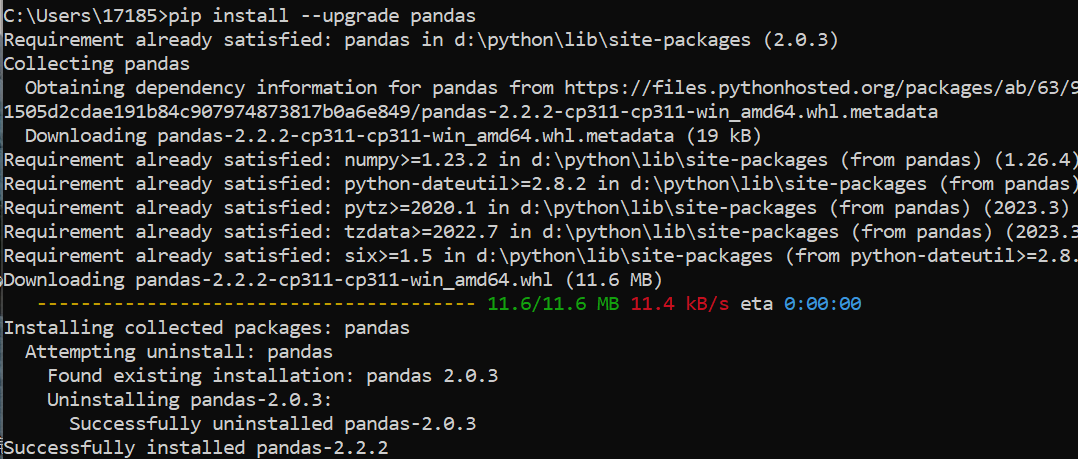


## 2. 软件包管理

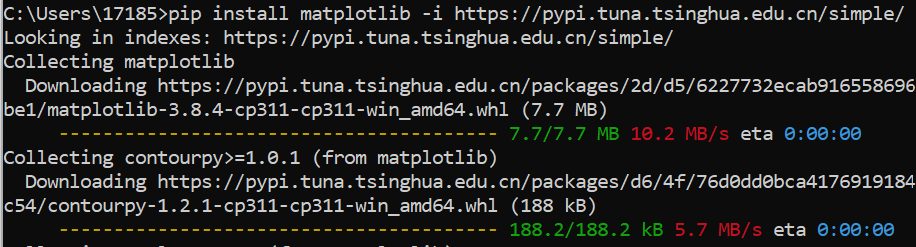
### （1）使用pip安装NumPy库



### （2）升级已安装的Pandas库

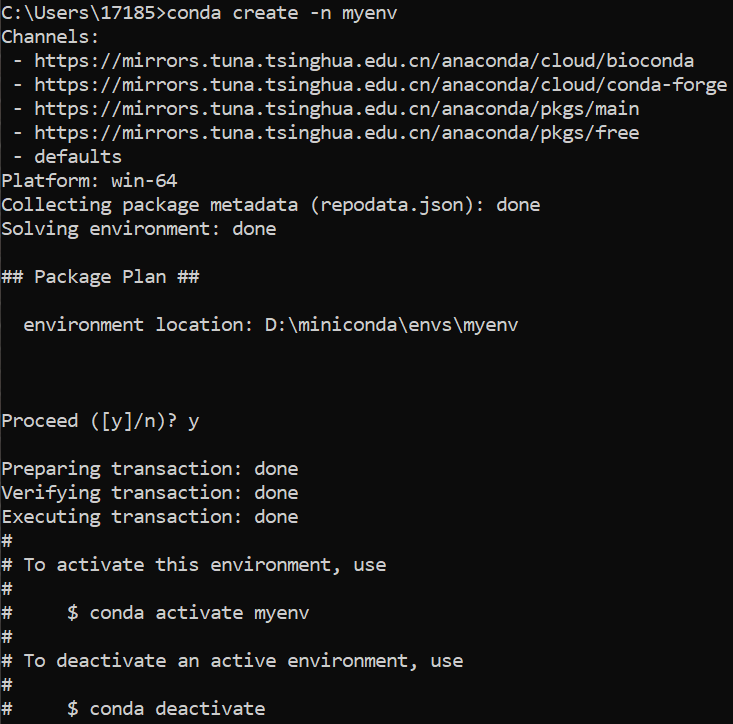


**（3）安装Matplotlib库**

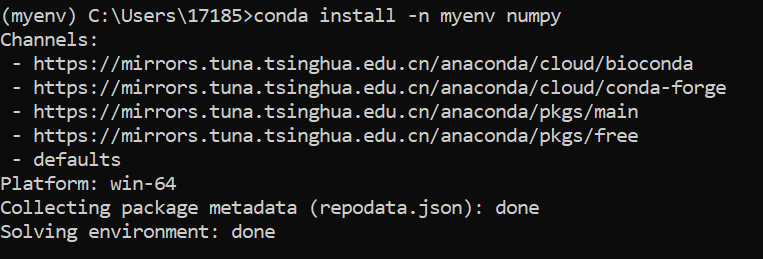


**（4）使用conda创建一个名为"myenv"的新环境，并在其中安装NumPy和Pandas库**

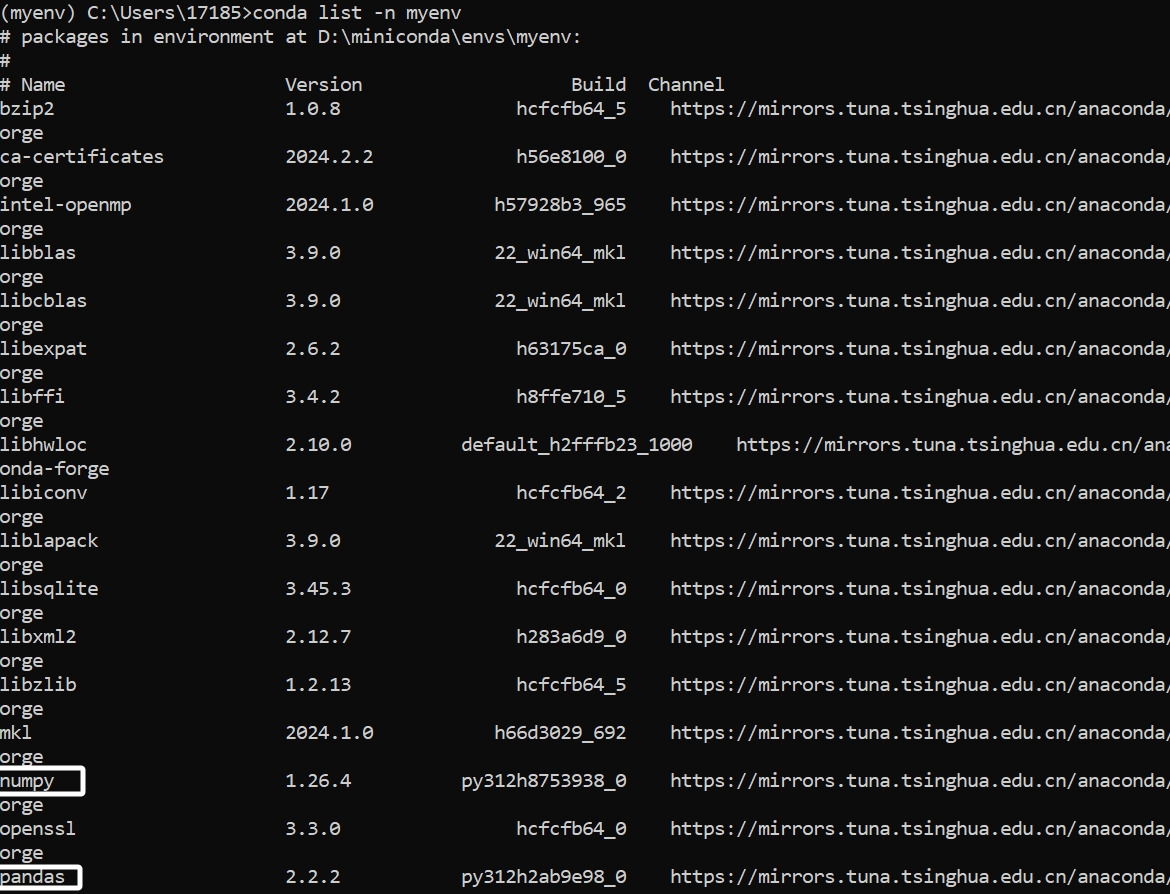
创建环境：



安装库：



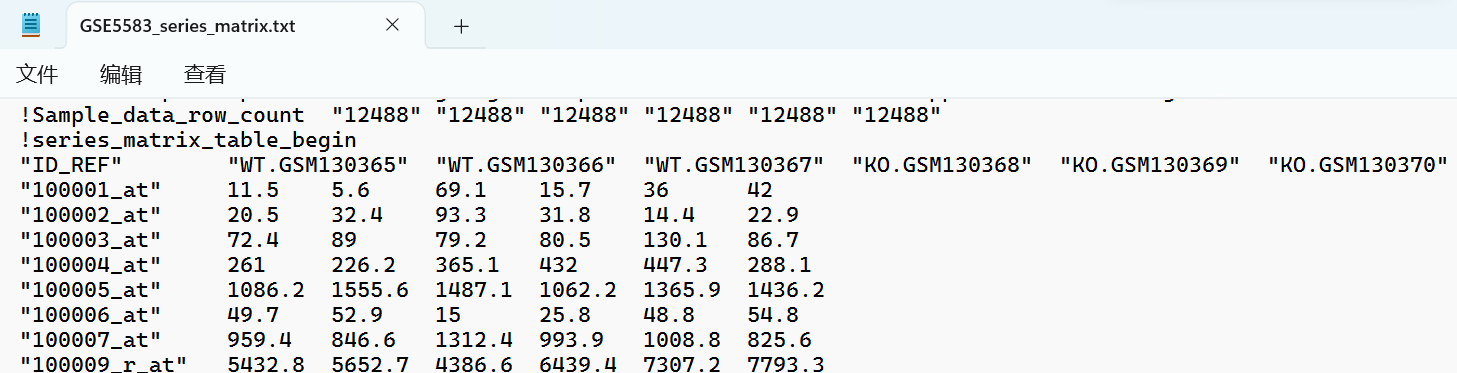
**（5）列出当前conda环境中已安装的所有软件包**



## 3. 生物信息分析

### （1）下载数据

下载得到的数据结构如下：



### （2）数据预处理及展示

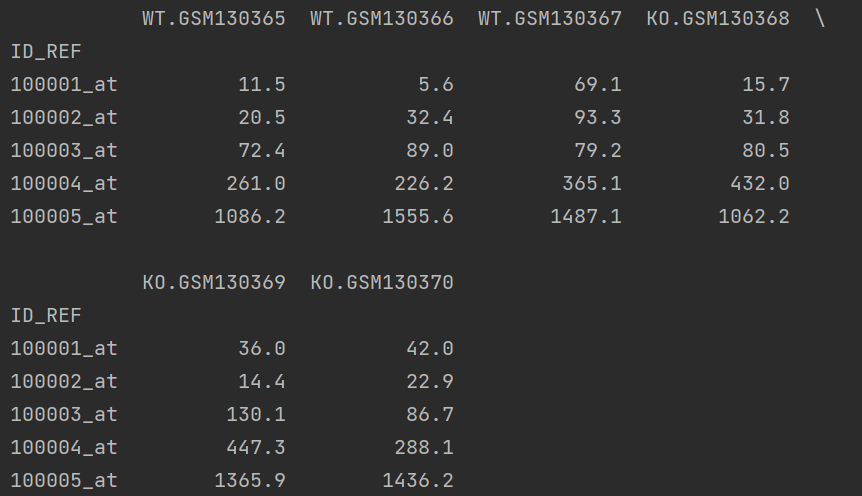
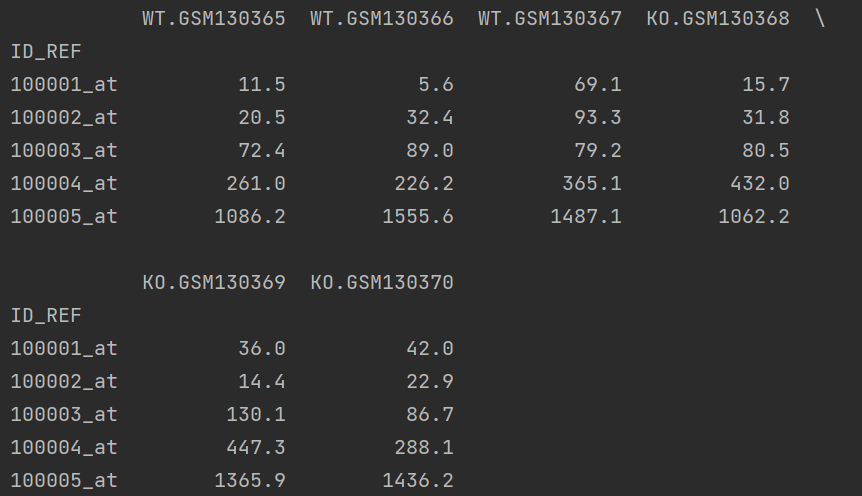
**基本思路：**

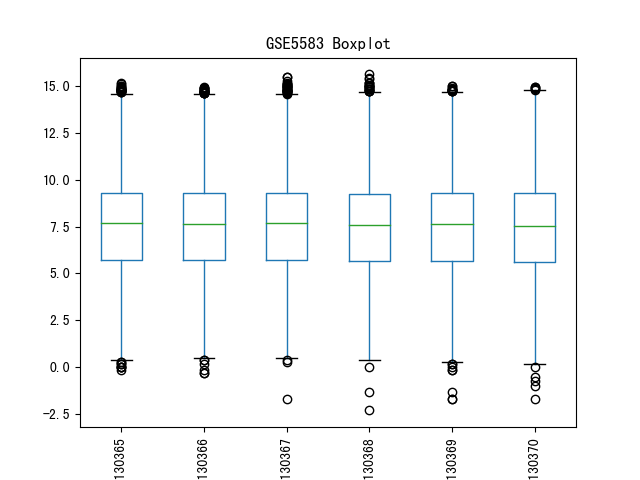
* 读取数据集，使用pandas包的read\_table函数
* 对数据进行log2标准化处理
* 绘制每个阵列的箱线图
* 绘制密度图，查看不同样本之间是否有总体差异
* 计算每个基因（行）wt和ko样本的表达平均值，并计算折叠变化fold change，绘制折叠变化的直方图

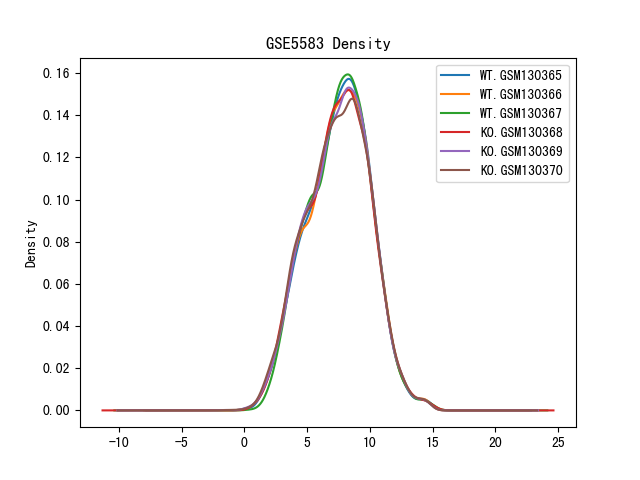
**实现代码：**

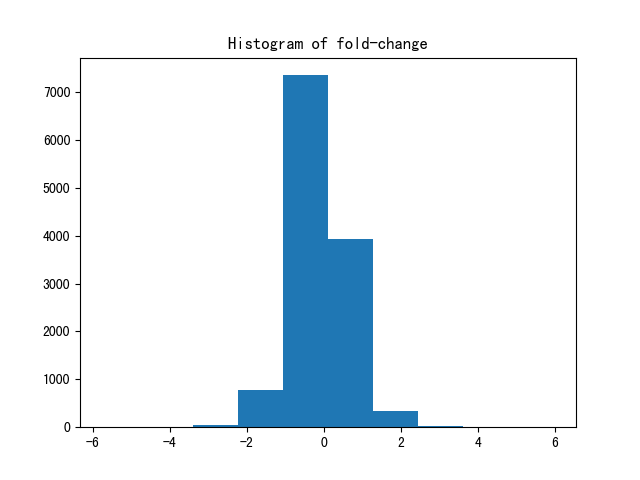


**运行结果：**



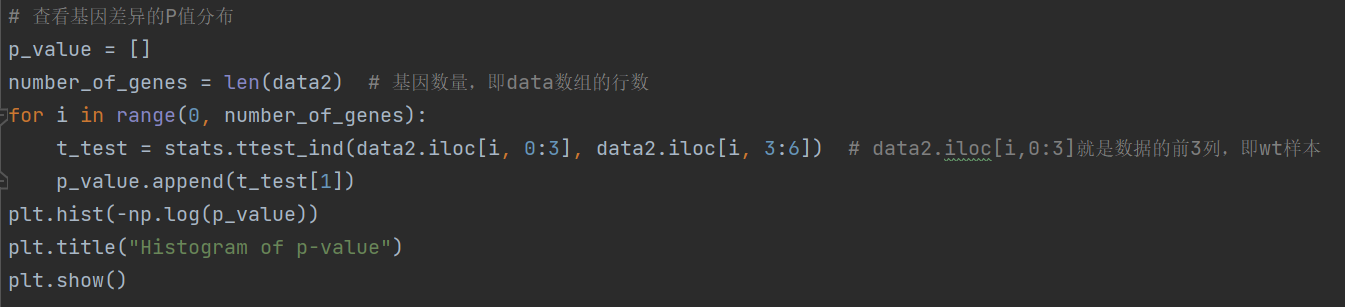




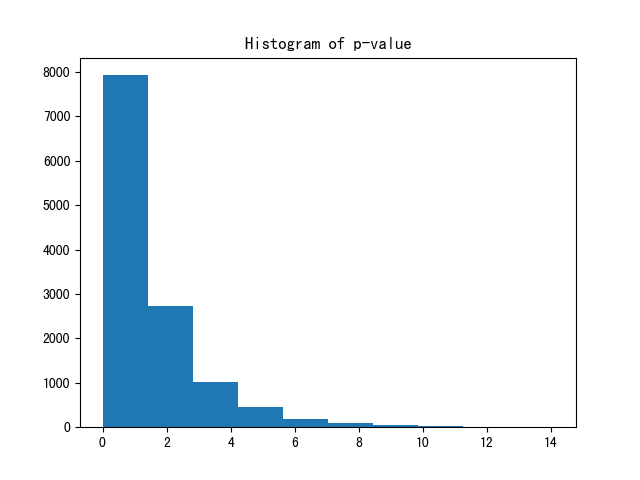
### （3）表达差异显著的基因

**基本思路：**使用t检验计算每个基因差异表达的p值，并绘制p值的直方图

**实现代码：**



**运行结果：**



# 第二次上机

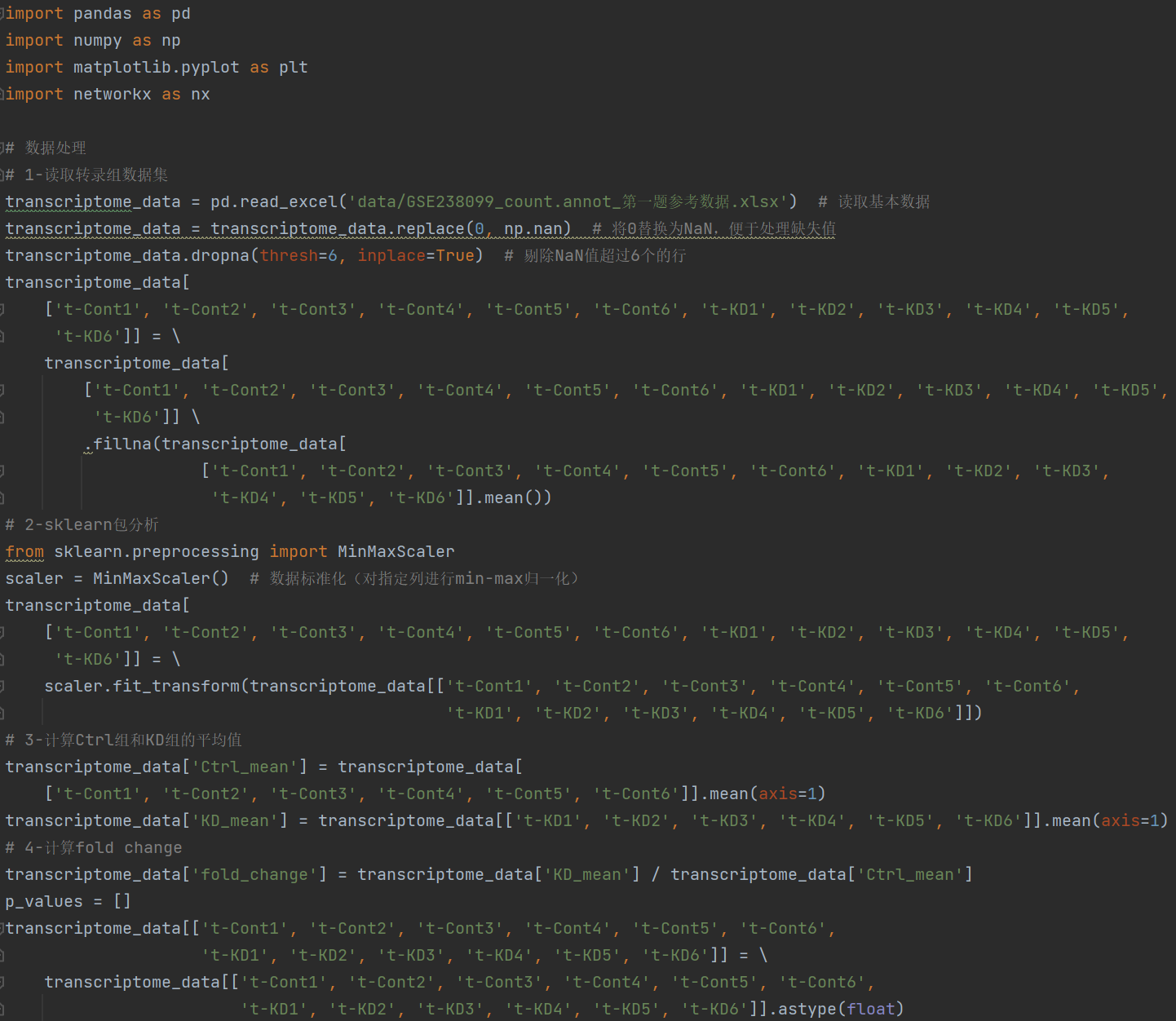
## 1. 生物信息分析

### （1）数据处理

**基本思路：**

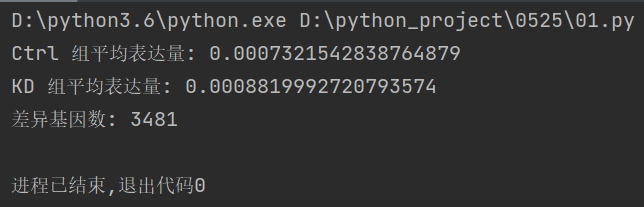
* 读取转录组数据集，使用pandas包的read\_excel函数
* 数据标准化，使用sklearn包中preprocessing模块内的MinMaxScaler函数，对指定列进行min-max归一化
* 计算Ctrl组和KD组的平均值以及fold change，并将结果保存到Ctrl\_mean、KD\_mean和fold change列中
* 使用scipy包中stats模块内的t函数计算p值，之后使用pandas包的apply函数进行计算并将结果保存到p\_value列中
* 根据基因的fold change筛选差异基因，注意要预先设置一个阈值
* 计算不同样本类型之间基因表达的差异，即求出Ctrl\_mean与KD\_mean列的平均值，以及差异基因数

**实现代码：**

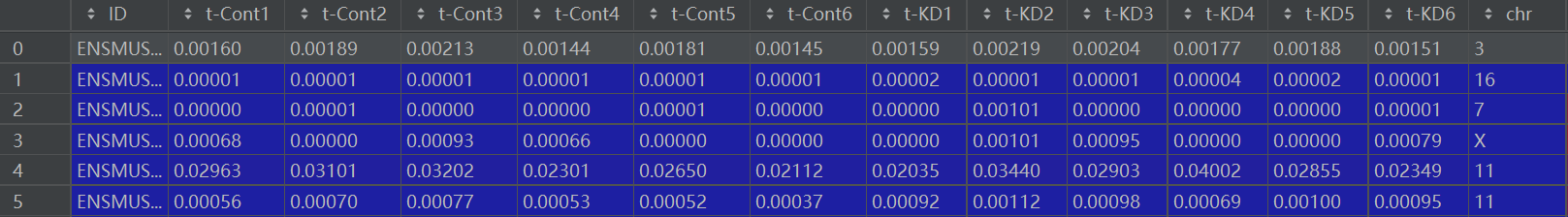


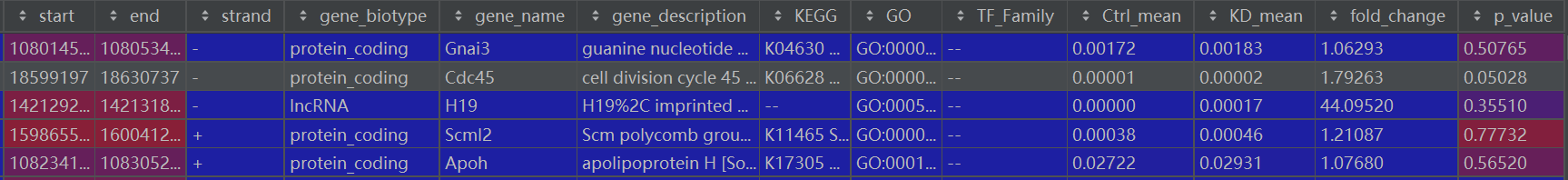


**运行结果：**



transcriptome\_data矩阵结构如下：





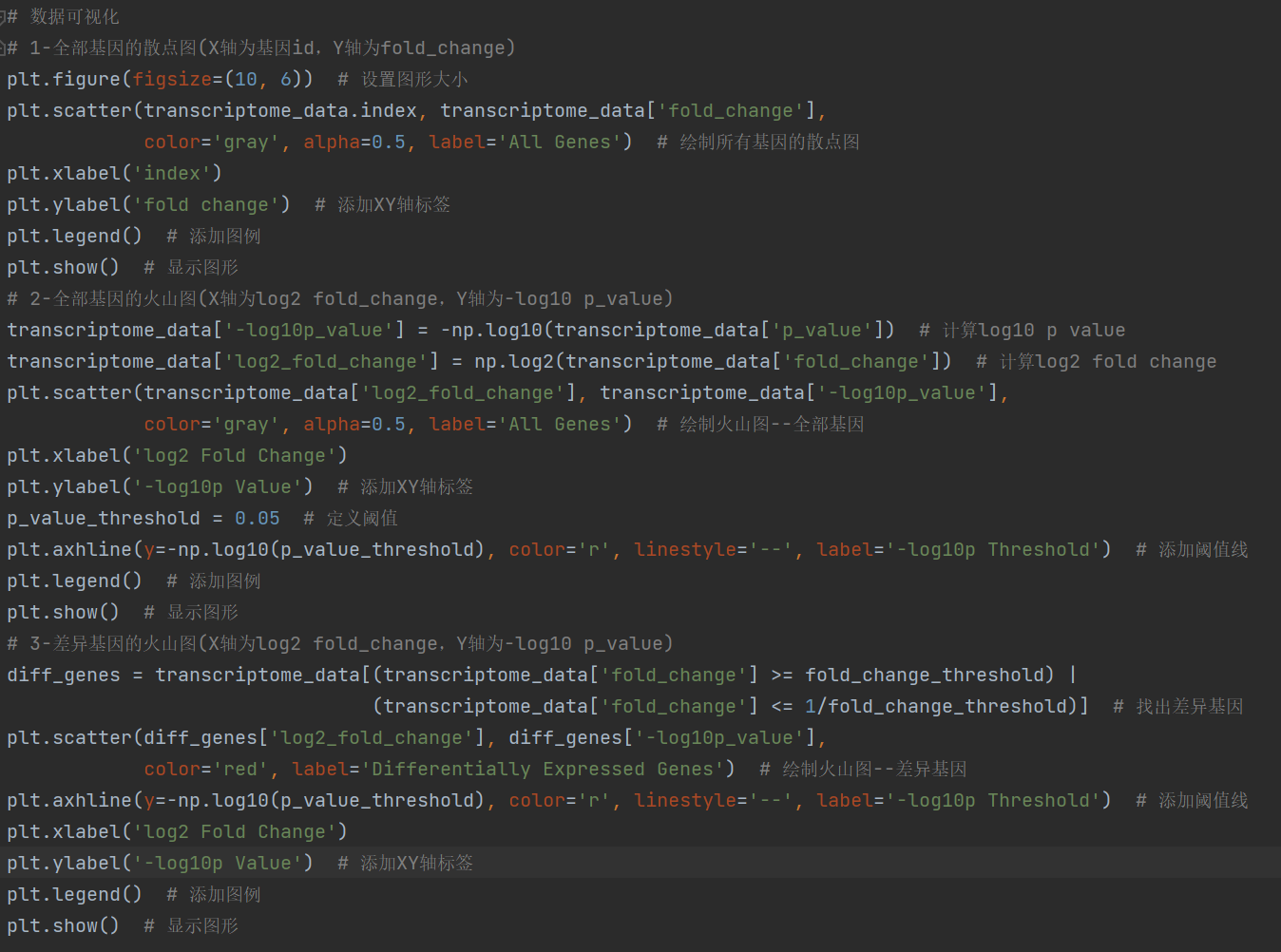
可以看到后面有刚刚计算出的Ctrl\_mean、KD\_mean、fold\_change和p\_value列

### （2）数据可视化

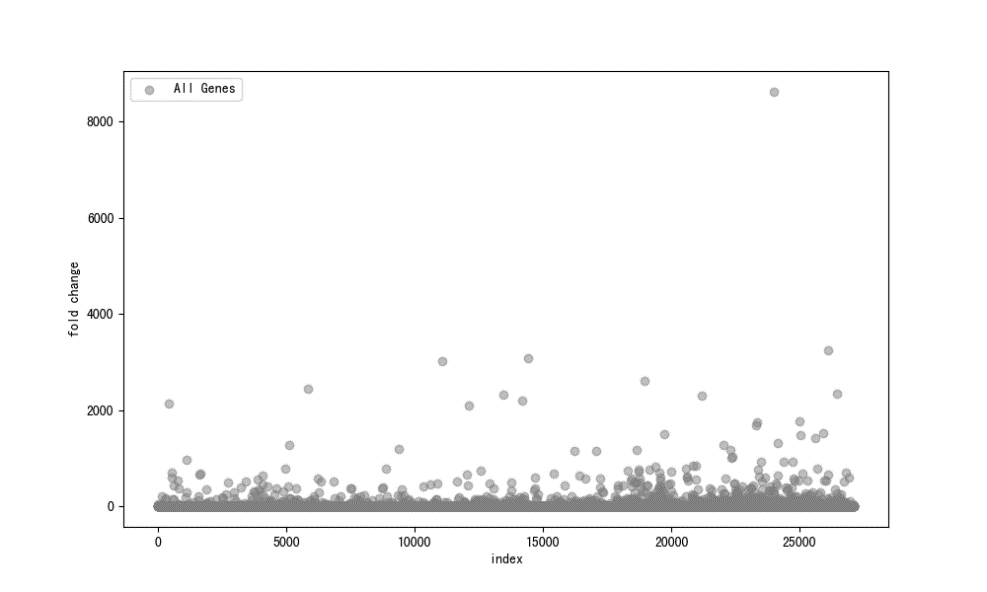
**基本思路：**

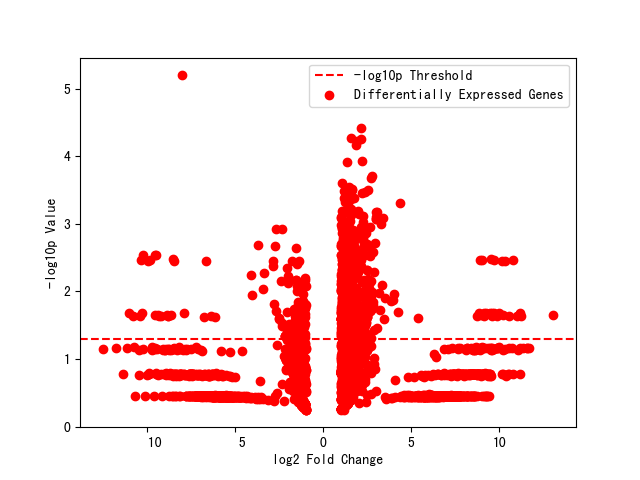
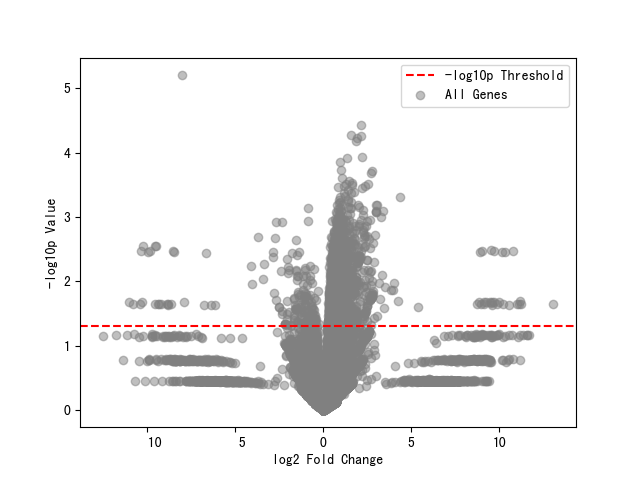
* 绘制全部基因的散点图，X轴为基因id（索引），Y轴为基因的fold\_change
* 绘制全部基因的火山图，需要先计算出每个基因的-log10(p\_value)以及log2(fold\_change)，X轴为基因的log2 fold\_change，Y轴为基因的-log10 p\_value
* 绘制差异基因的火山图，需要先根据自设的阈值以及每个基因的fold change找出差异基因，X轴为基因的log2 fold\_change，Y轴为基因的-log10 p\_value

**实现代码：**



**运行结果：**





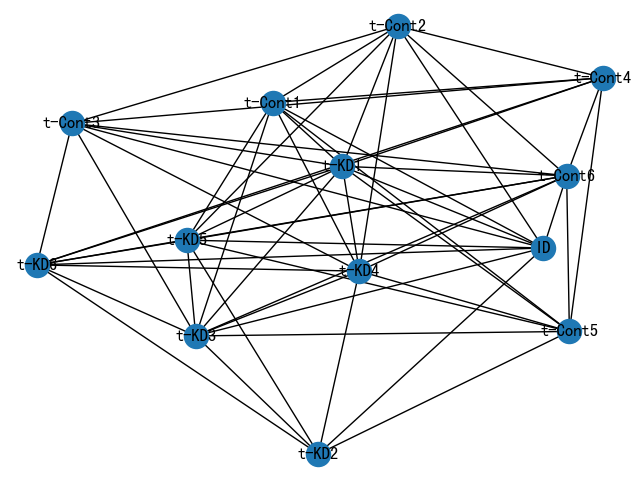
### （3）基因关联网络构建

**基本思路：**需要有基因间的相关性数据（这里我直接使用随机数生成了），使用networkx包构建网络——首先添加节点，之后根据自设阈值选择相关性大于某个值的边进行添加，最后就可以绘制了。

**实现代码：**



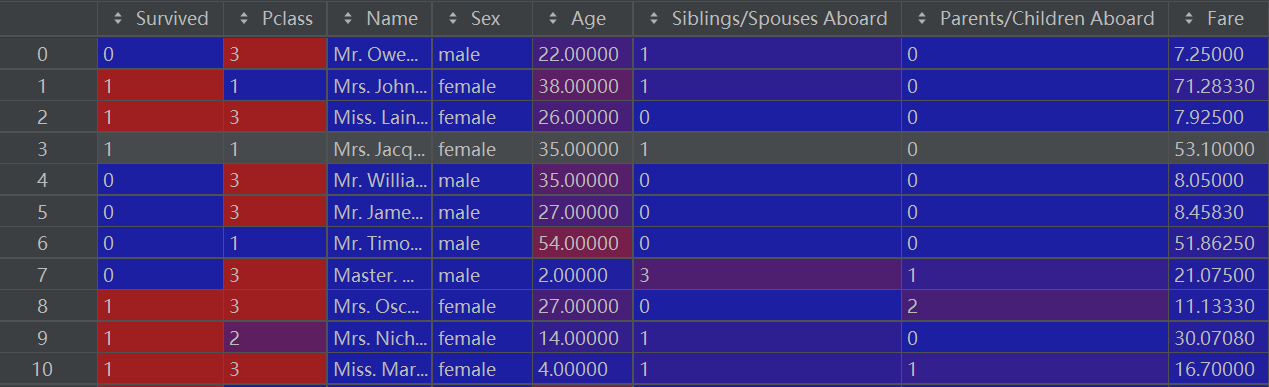
**运行结果：**

****

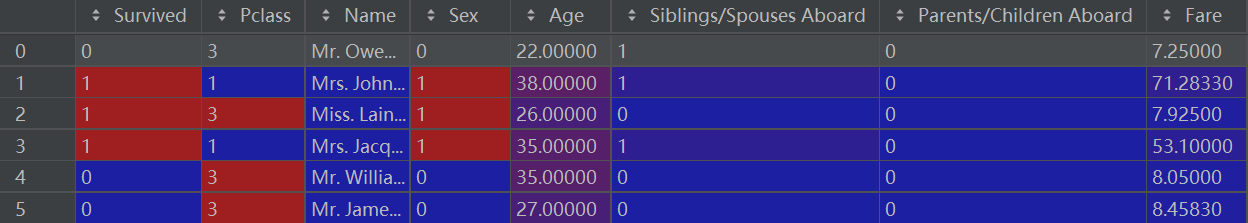
## 2. 机器学习

**基本思路：**

* 读取数据：使用pandas包的read\_csv函数

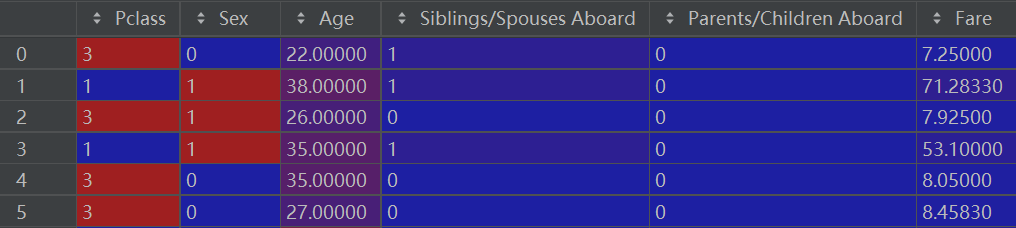


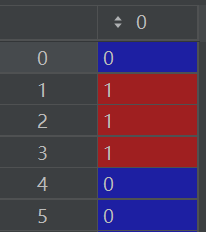
* 数据预处理：处理缺失值，转换类别变量为数值变量等。因为我这里使用的sklearn包中fit函数必须要求所有列均为数值型，因此该数据集内需要把Sex列转成数值型，我设定将male->0、female->1



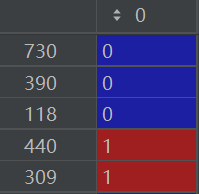
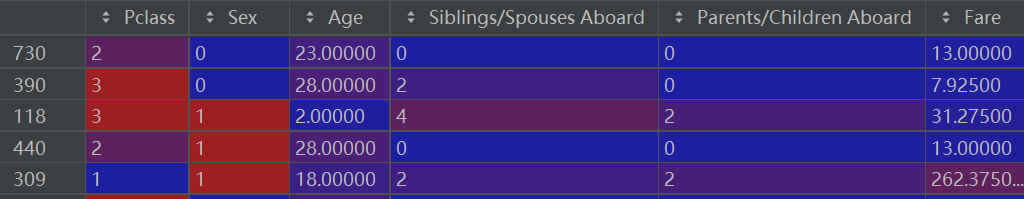
* 划分特征和目标变量：Y为是否survived（用0和1表示），X为除去survived和name的其它列；并划分数据集为训练集(X\_train, y\_train)和测试集(X\_test, y\_test)

X：

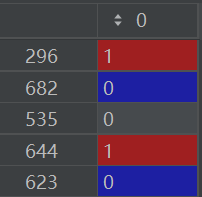
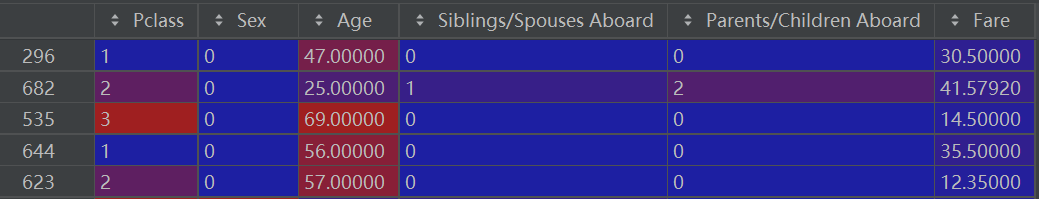
Y：



X\_train, y\_train：



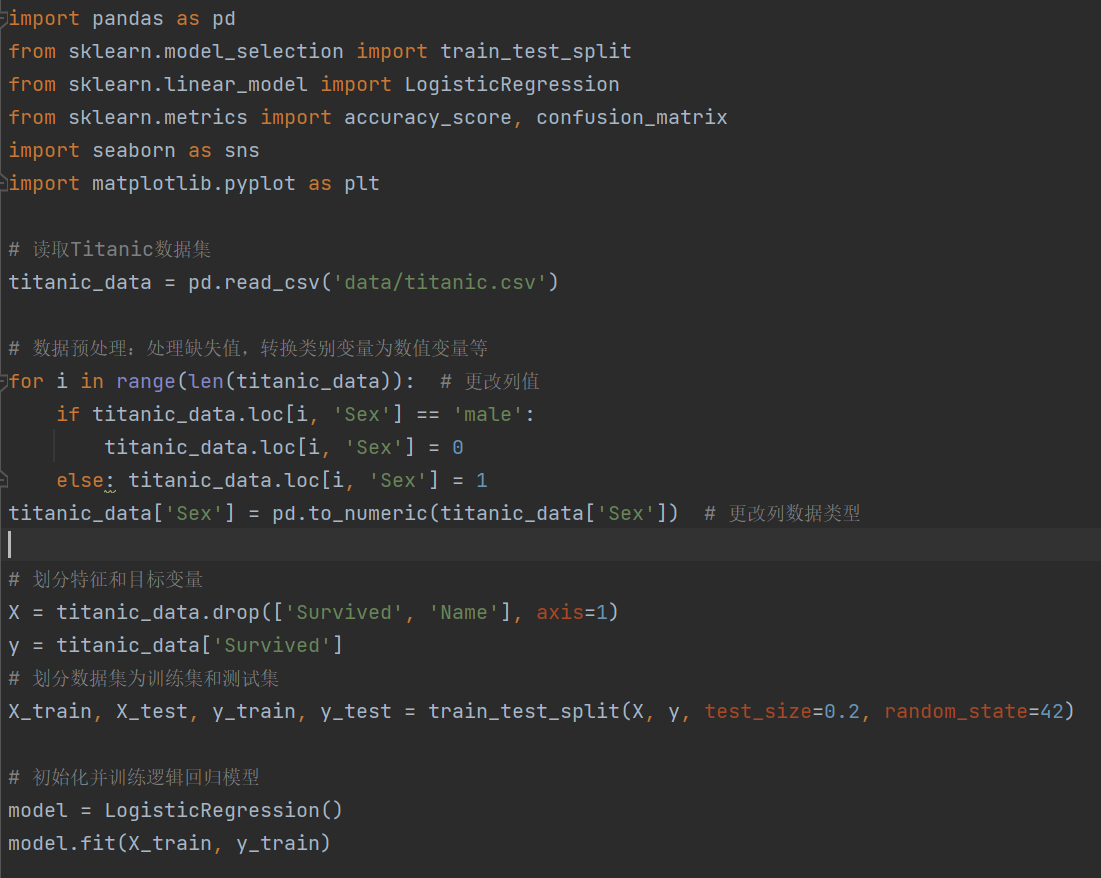
X\_test, y\_test：

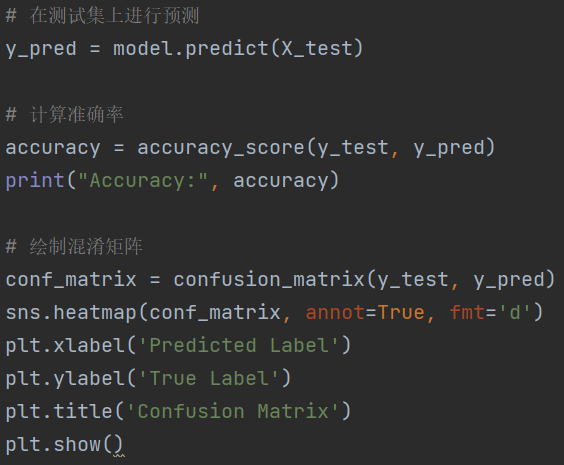


可以看到训练集和测试集中X-y的索引是分别相互对应的

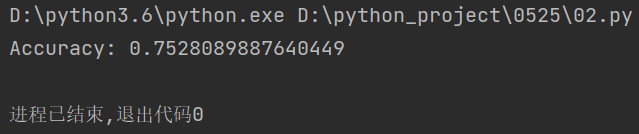
* 初始化并训练逻辑回归模型：使用LogisticRegression函数创建一个model，并使用fit函数进行训练
* 在测试集上进行预测：使用predict函数
* 计算准确率：使用sklearn包中metrics模块内的accuracy\_score函数
* 绘制混淆矩阵：使用sklearn包中metrics模块内的confusion\_matrix函数获得数据，使用seaborn包内的heatmap函数进行绘制

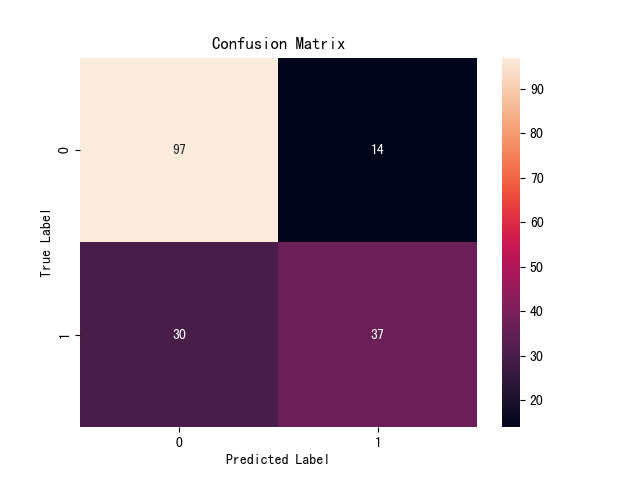
**实现代码：**





**运行结果：**

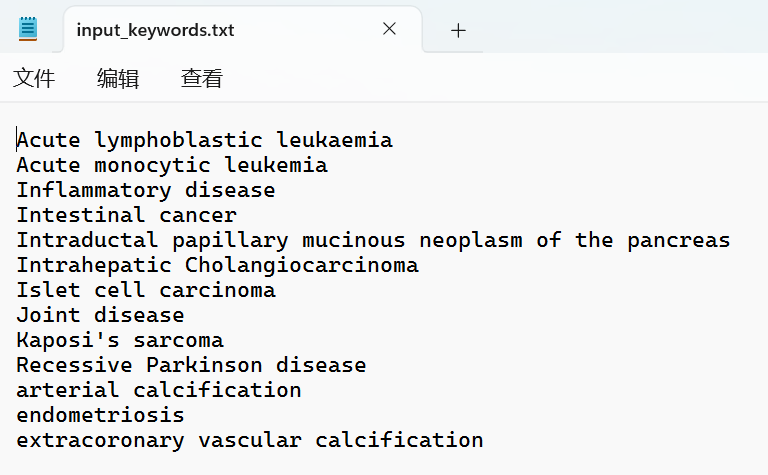




## 3. 数据获取

**基本思路：**

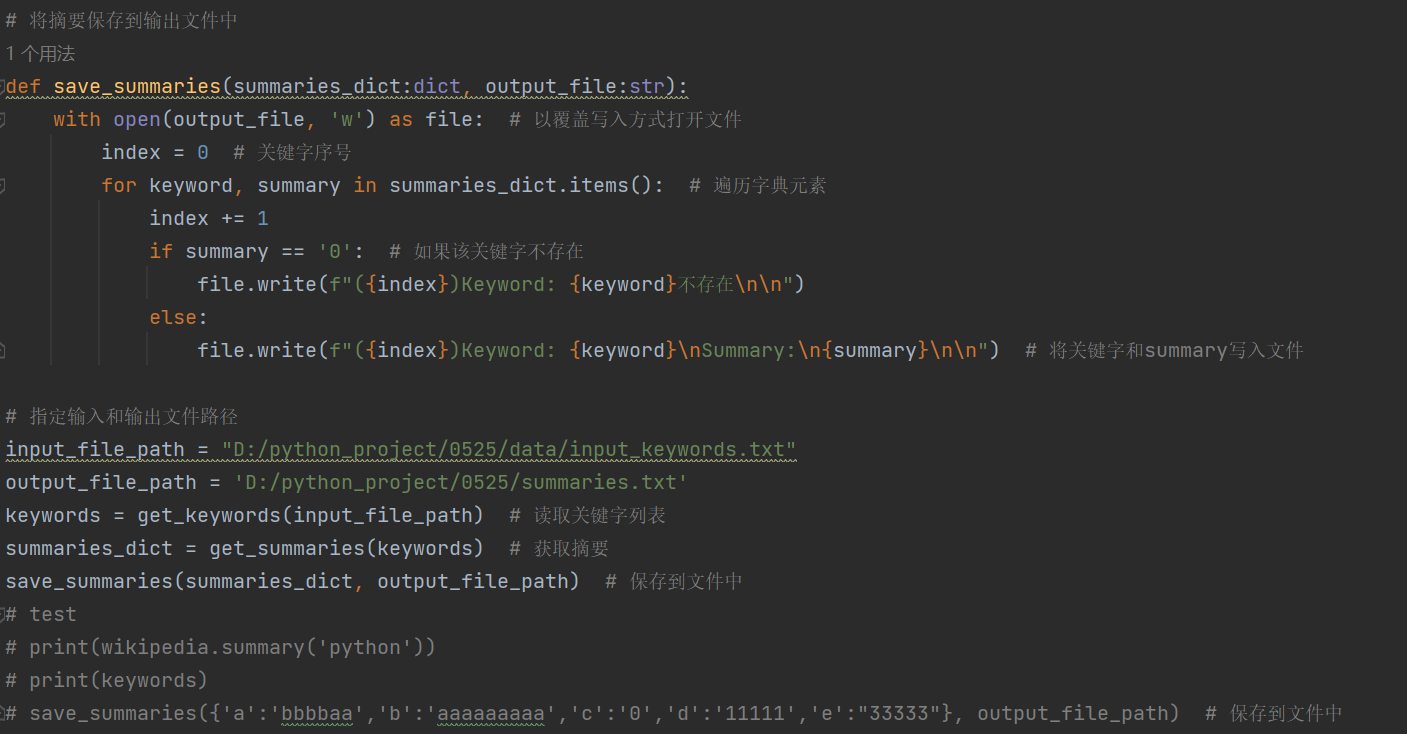
* 从指定的txt文件中读取关键字列表keywords（由于网络不稳定等原因，需减少运行时间，因此我减少了keywords的数量）



* 获取关键字的summary：使用wikipedia包中的summary函数，并以字典形式保存（键为keyword，值为它的summary）
* 将结果保存到输出文件中

**实现代码：**





**运行结果：**（截取部分）

