1. 数据填充

（1）表单一填充（热卡填充，众数填充）

首先对数据缺失处进行分析，为A铅钡风化与C铅钡风化。

填充思路：热卡填充，即寻找同类中化学成分与待填充文物相关性最大的数据进行填充颜色。（属性值填充）对数据分析过程中按照权值加和不同文物部位的化学成分值，得到单一文物的化学成分值。对于C类文物大部分采样点为无风华点对于风华文物缺乏参考意义，故采用众数填充。

1. 表单二，三对0值的填充（最小值\*sigma填充，最小化协方差矩阵的扭曲）

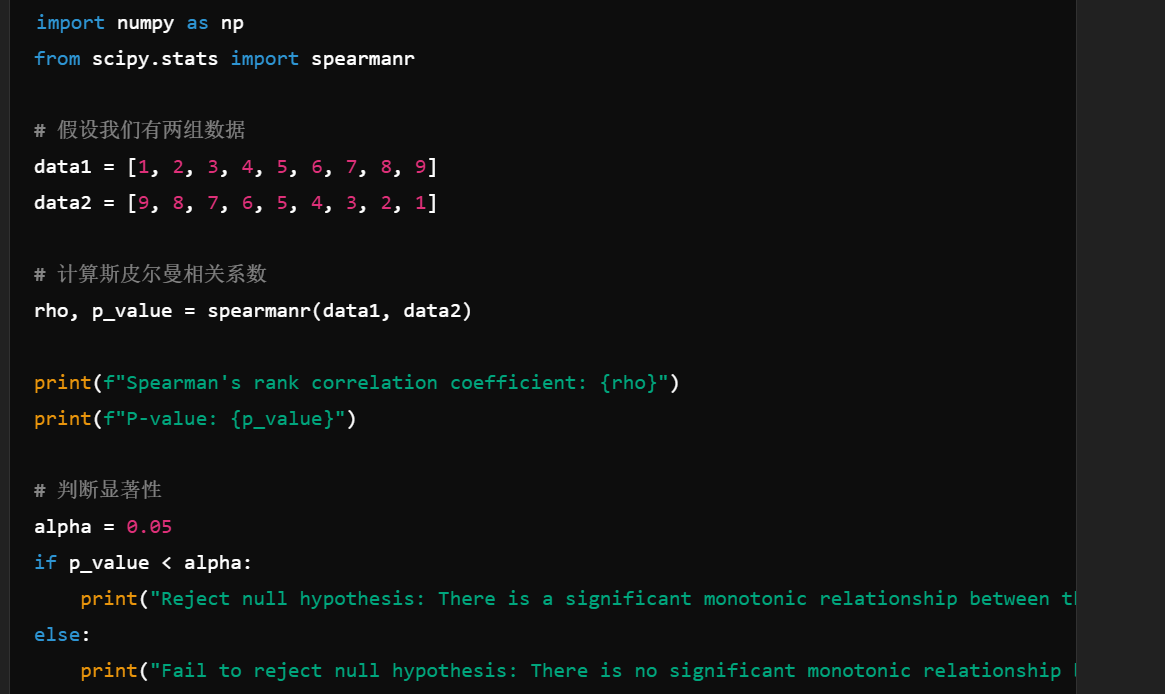
# 期望协方差矩阵（理论矩阵） Sigma\_expected = np.array([[0.04, 0.01], [0.01, 0.09]])

# 实际协方差矩阵（观测矩阵） Sigma\_actual = np.array([[0.05, 0.02], [0.02, 0.08]])

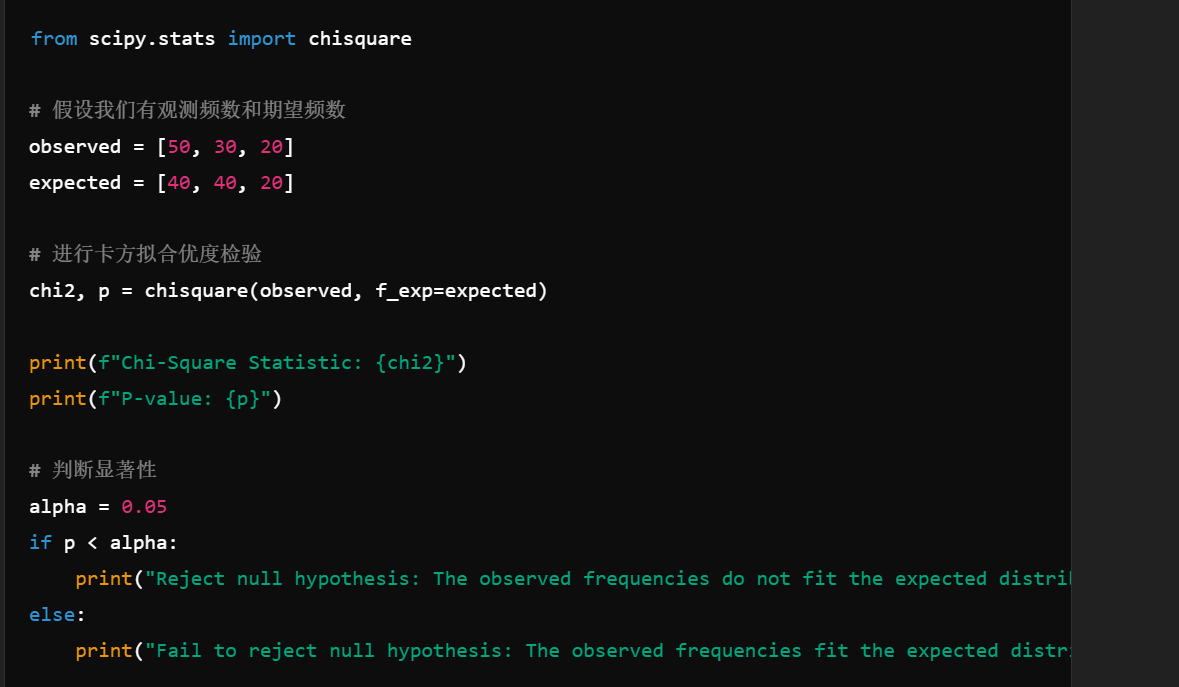
# 计算相对误差矩阵 R = (Sigma\_actual - Sigma\_expected) / Sigma\_expected

# 计算均方根误差 mse = np.sqrt(np.mean(R\*\*2))

1. 相关性分析（表单一）
2. spearman相关系数

(2)基于风化与未风化的各指标频数统计图

(3)卡方检验（检验分类变量之间是否存在关系）



（4）针对表单二采用环形图表进行分析

1. 分析分类规律

结果划分：【1】.

clr变换（常用于化学成比例分析）

决策树确定指标（铅钡），利用指标进行聚类

【2】偏最小二乘判别分析：

原理为从特征值中提取主成分（本例提取的主成分数为2），将y值转化为向量，而后进行线性回归。回归将得到一个线性方程，通过设置阈值将预测值分类。

检验方法：

（1）在坐标图中构建点位位置，图示检验亚类之间的分离性

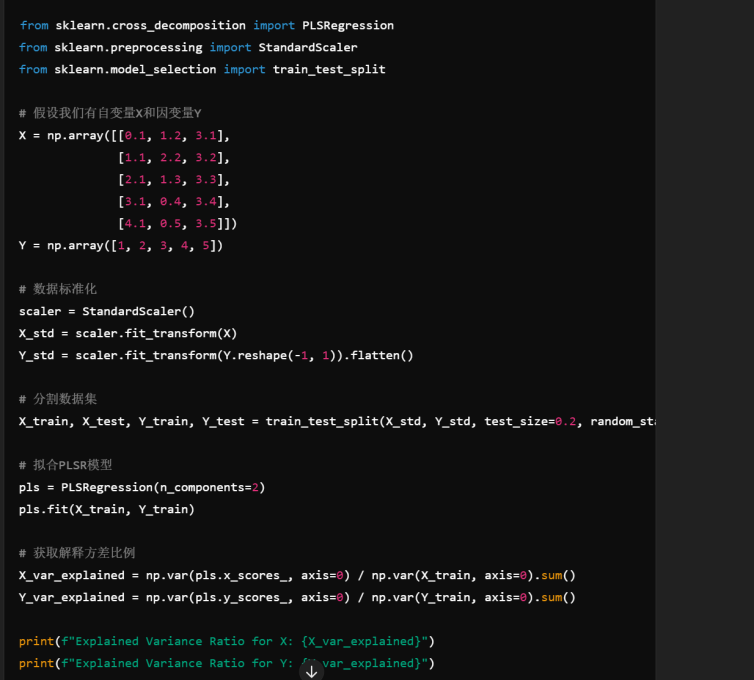
（2）通过构建偏最小二乘线性方程与坐标轴的交点（b0）来说明结果的合理性（个人感觉无法充分说明模型合理性，更应该通过测试集上的预测准确率去检验模型）

4.预测分类结果

【1】决策树选取决策变量

【2】偏最小二乘预测

1. 选取合适的主成分（如下代码，比较各成分的贡献），同时进行交叉检验使误差最小化。（选取的主成分来自上一步VIP检验VIP值最大的四个成分）

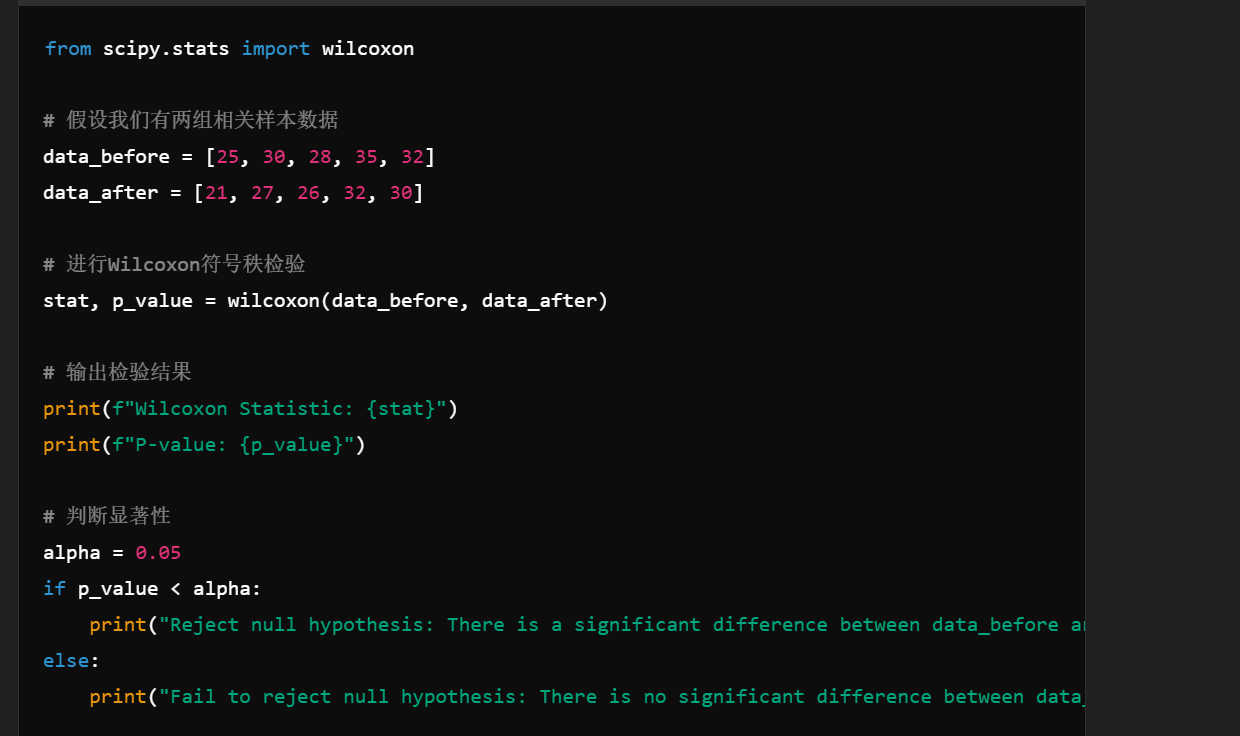


1. 进行预测分类
2. 化学成分相关性分析
3. 同一类别分析



计算相关性并绘制热力图（基于相关矩阵）

1. 不同类别分析



Wilcoxon检验两样本数据之间的差异

偏最小二乘用法总结：

1. 找到模型贡献值最大的成分
2. 检验模型的分类合理性