五一数学建模竞赛

承　诺　书

我们仔细阅读了五一数学建模竞赛的竞赛规则。

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与本队以外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的，如果引用别人的成果或其它公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们愿意承担由此引起的一切后果。

我们授权五一数学建模竞赛组委会，可将我们的论文以任何形式进行公开展示（包括进行网上公示，在书籍、期刊和其他媒体进行正式或非正式发表等）。

参赛题号（从 A/B/C 中选择一项填写）：_____

参赛队号：_____

参赛组别（研究生、本科、专科、高中）：_____

所属学校（学校全称）：_____

参赛队员：　队员 1 姓名：_____

　　队员 2 姓名：_____

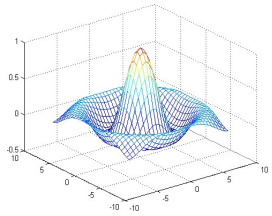
　　队员 3 姓名：_____

联系方式：　Email：_____　联系电话：_____

日期：　_____年___月___日

（除本页外不允许出现学校及个人信息）

五 一 数 学 建 模 竞 赛



题 目：_____

关键词：

摘　要：

一、 问题重述

1.1 问题背景

在中国古代，丝绸之路作为中西方文化交流的主要渠道，玻璃作为当时最具代表性的身份象征，体现其拥有者的尊贵。早期玻璃是从西亚和埃及地区常被制作成珠形饰品传入我国。我国古代大能通过研究玻璃的成分以及其原理在本土就地取材制作，虽然外观相似，但是化学成分却不相同。

玻璃的主要原料是石英砂，主要化学成分是二氧化硅 (SiO2）。在煅烧玻璃时，需要添加一些辅助材料作为助熔剂，因此会产生很多其他的化学成分。石灰石煅烧之后会转化为氧化钙（CaO）。铅钡玻璃在烧制的过程中加入铅矿石作为助熔剂，其中氧化铅（PbO）、氧化钡（BaO）的含量较高，通常被认为是我们自己发明的玻璃品种，楚文化的玻璃就是以铅钡玻璃为主。其中，高钾玻璃是以含钾量高的物质如草木灰作为助熔剂烧制而成的。

1.2 问题的提出

由于古代的保存手段比较低端，因此古代玻璃极易受埋藏环境的影响而风化。在风化过程中，内部元素与环境元素进行大量交换，导致其成分比例发生变化，从而影响对其类别的正确判断。因此，需要通过数字数据来对玻璃的风化状态以及化学成分进行分析。

1.3 问题的分解

1.3.1 问题一分解

1. 分析玻璃文物的表面风化与玻璃类型、纹饰和颜色之间的关系；
2. 分析文物样品表面有无风化化学成分含量的统计规律；
3. 预测风化点检测数据在风化前的化学成分含量；

1.3.2 问题二分解

1. 分析高钾玻璃、铅钡玻璃的分类规律；
2. 对每个玻璃类别进行亚分类，给出划分方法和结果；
3. 分析分类结果的合理性和敏感性；

1.3.3 问题三分解

1. 对附件表单 3 中的未知类别玻璃文物进行鉴别；
2. 对分类结果敏感性进行分析；

1.3.4 问题四分解

1. 分析不同类别玻璃文物中化学成分之间的关系；
2. 比较不同类别玻璃文物中化学成分的差异性；

二、 问题分析

2.1 数据集分析

样本数据的采集方式有很多种，不同的方式会得到不同质量的样本数据，一般情况下我们获取的原始数据并不能直接为我们所用, 需要根据我们所分析的问题通过预处理的方法将数据转变成合适的干净数据。数据预处理一般包含数据清洗、数据的缺失值和异常值处理、特征工程等几个步骤。本文中涉及到的数据处理、分

析以及模型的建立与评估均采用 Python3.7 版本来实现，其中主要使用 Python 中的 Pandas、Numpy、Seaborn、Matplotlib、Sklearn 等库[6]。

本文主要通过 SPSSPRO 软件中的 Notebook 应用来建立 jupyter 文件，然后读取给定的数据集，导入建模所需的工具库。

2.2 缺失值处理

对于缺失值的处理有三种方法，分别是缺失值插补法、删除样本法与直接使用缺失值的特征。插补法是根据样本数据情况使用均值、众数或者相邻样本来补充缺失值; 删除样本法是对于样本数据的变量缺失过多或剩余变量无明显特征，无法体现出样本的特征信息，删除该样本对整体结构无法造成影响; 直接使用缺失值特征是在某些特殊情况下，将缺失值映射为一个类别特征，则可不对其进行处理。

通过 isnull() 函数查看数据集是否包含缺失值以及异常值，结果如表1。

表 **1**　缺失值检测

特证名	空值个数
文物编号	0
纹饰	0
类型	0
颜色	4
表面风化	0

2.2.1 问题一分析

第一小问需要进行差异性分析：由于所有特征变量为定类变量，因此进行卡方检验分析确定自变量与因变量之间的关系。通过 SPSSPRO 进行求解。分析显著性 p 值是否小于 0.05 来分析其差异性关系。

第二小问是变化规律分析：通过计算样本数据的均值、计数、中位数、标准差、最大值、最小值和求和共七个统计量进行描述性统计分析、散点图统计分析、正态分布检验直方图等可视化展现等，总结变化情况。

第三小问是预测化学成分，根据风化前后的数据规律，总结各个化学成分的变化情况，找出之间的关系并预测风化前的含量。

2.2.2 问题二分析

第一小问需要针对高钾玻璃和铅钡玻璃不同化学成分的数值进行统计，找到分类的依据。

第二小问需要在第一小问的基础上进行亚类划分，观察化学成分在风化前后的颜色变化、纹理变化等，并给出相关的分类依据。

第三小问需要对数据进行灵敏性检验，并给出合理性依据。

2.2.3 问题三分析

第一小问需要将附件表单三中附件中有无风化的情况进行分类讨论，结合问题 2 中模型的结论，对表单三中不同类型的玻璃进行分类。

第二小问通过对某一类化学成分进行增加或减少，观察分类情况是否发生变化，给出模型的稳定性结论。

2.2.4 问题四分析

第一小问需要选取具有代表性的化学成分进行灰色关联分析，与问题一第二小问的区别在于少了一个有无风化的条件，选择化学成分占比最大的作为因变量，其余作为自变量，建立灰色关联分析模型，计算其灰色关联度的情况。

第二小问需要对不同类别的玻璃进行显著性检验，观察两种玻璃之间的差异性。

三、 模型假设

针对本文提交的问题，需要做如下模型假设：

1. 给定的数据真实有效，不包含造假成分；
2. 附件表单二和三中不含有成分表示含量为 0；
3. 对附件中的数据分析不考虑时间因素的影响；

四、 符号说明

符号	含义
Q	误差平方和
$\hat{b}_i$	模型拟合值
s^2	方差
$\bar{b}$	残差平方和
R	复相关系数
μ	位置参数
x	特征变量
Y	目标变量
θ	权值向量
b	偏置
$\hat{\theta}$	θ 的估计值

五、 问题求解

5.1 问题一的建模与求解

首先需要对玻璃表面风化情况与玻璃类型，纹饰和颜色的差异性进行分析，并结合玻璃的类型分析化学成分含量的变化规律以及预测风化前的化学成分含量，共需解决三个小问题，问题一建模分析流程图如下图1所示。

5.1.1 数据预处理

1. 首先进行数据预处理工作，根据题目要求: 将成分比例累加和介于 85% 105% 之间的数据视为有效数据。
2. 附件表单 1 中颜色列中的数据中，我们发现有四个空值，这里直接删去包含缺失值的样本数据。

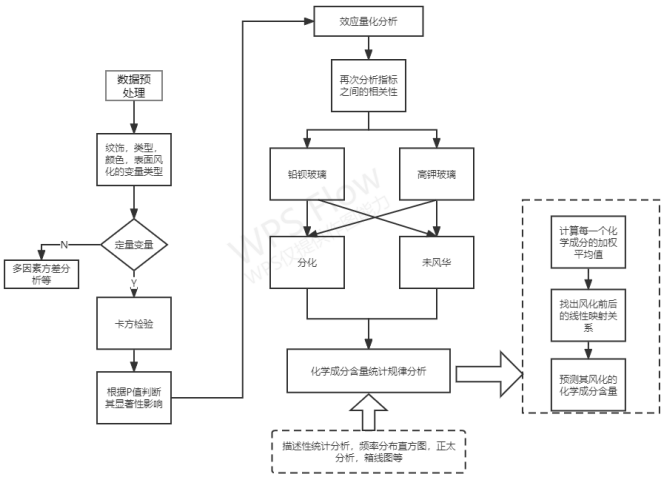


图 **1**　组织结构图

5.2 独热编码

独热编码（One-Hot Encoding），又称之为一位有效编码，具体是使用 N 个寄存器对应 N 个状态，并且相应状态都有它独立的寄存器位，并且在任意时候，其中只有一位有效。即，只有一位是 1，其余都是零值。独热编码是利用 0 和 1 表示一些参数，使用 N 位状态寄存器来对 N 个状态进行编码。

通过分析数据，发现玻璃类型、纹饰、颜色变量都与表面是否风化具有一定的相关性。因此通过玻璃类型与表面分化、纹饰与表面风化、颜色与表面分化两两进行比较分析，得出 spearman 相关系数和复相关系数数据，发现 spearman 相关系数精确度较高，因此使用 spearman 相关系数方法来分析此问题，热力图如图2所示：

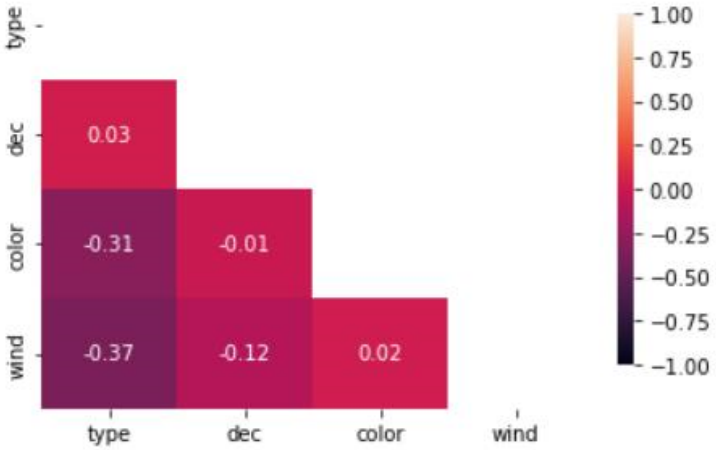


图 **2**　复相关热力图系数

5.2.1 针对表面风化情况做卡方检验

首先通过 Python 代码将附件表单一和表单二中的数据进行合并，方便接下来的统计，通过观察数据发现纹饰、类型、颜色、表面风化均为定类变量，针对多组定类变量之间的差异性分析我们可以采用卡方检验。

变量 X：表面风化；变量 Y：纹饰，类型，颜色，使用 SPSSPRO 软件进行交互分析，得出如下图4所示的卡方检验表。

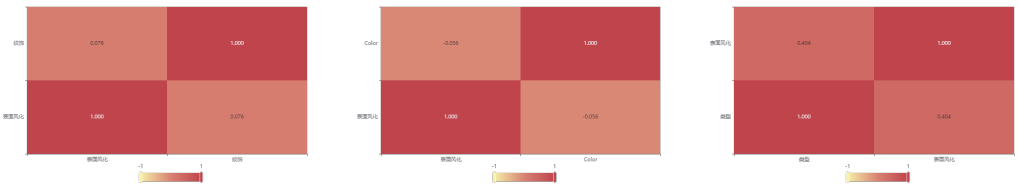


图 **3**　**spearmand** 系数对比图

题目	名称	表面风化		总计	X^2	校正X^2	P
		无风化	风化				
纹饰	A	11	9	20	5.747	5.747	0.056*
	B	0	6	6			
	C	13	15	28			
类型	铅钡	12	24	36	5.400	4.134	0.020**
	高钾	12	6	18			
颜色	浅绿	2	1	3	6.287	6.287	0.507
	浅蓝	8	12	20			
	深绿	3	4	7			
	深蓝	2	0	2			
	紫	2	2	4			
	绿	1	0	1			
	蓝绿	6	9	15			
	黑	0	2	2			

图 **4**　类型-颜色-表面风化卡方检验表

图表说明：上表展示了模型检验的结果，包括数据的频数、频数百分比、卡方值、显著性 P 值。

1. 分析模型是否呈现出显著性（P 值小于 0.05 或 0.01）；
2. 若呈现显著性，拒绝原假设，则说明各样本之间存在显著性差异。具体根据类别的差异百分比进行描述。反之数据不存在显著性差异。

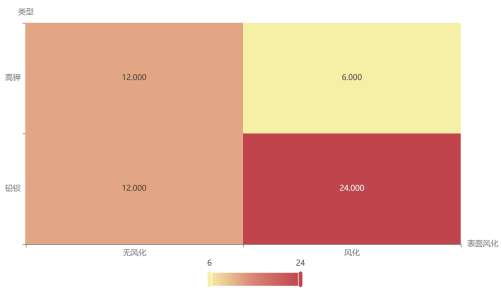


图 **5**　表面风化-类型热力图

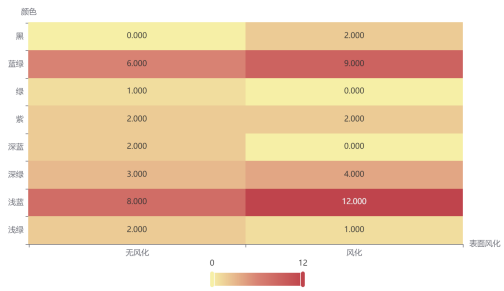


图 **6**　表面风化-颜色热力图

在此基础上进行效应量化分析，包括 phi、Crammer's V、列联系数、lambda , 用于分析表面风化与其余三个指标的相关程度，量化分析指标解释如下:

1. phi 系数: phi 相关系数的大小，表示两样本之间的关联程度。当 phi 系数小于 0.3 时，表示相关较弱; 当 phi 系数大于 0.6 时，表示相关较强
2. Cramer's V: 与 phi 系数作用相似，但 Cramer's V 系数的作用范围较广。e3）列联系数: 简称 C 系数。

3. lambda: 用于反应自变量对因变量的预测效果

使用 SPSSPRO 进行操作，得出结果如下表2所示。

表 **2**　表面风化效应量化分析

字段名/分析项	Phi	Crammer's	列联系数	lambda
纹饰	0.326	0.326	0.310	0.000
类型	0.316	0.316	0.302	0.000
颜色	0.341	0.341	0.323	0.000

效应量化分析的结果显示，分析项：纹饰 Cramer’s V 值为 0.326，因此纹饰和表面风化的差异程度为中等程度差异；类型 Cramer’s V 值为 0.316，因此类型和表面风化的差异程度为中等程度差异；颜色 Cramer’s V 值为 0.341，因此颜色和表面风化的差异程度为中等程度差异。

5.2.2 不同类型玻璃表面有无风化化学成分统计规律

首先使用 SPSSPRO 软件针对描述性铅钡和高钾两种玻璃风化前后化学成分含量统计分析，结果如下表3所示：

表 **3**　高钾玻璃表面风化效应量化分析

变量名	风化情况	样本量	最大值	最小值	平均值	标准差	中位数	方差	峰度	偏度	变异系数 (CV)
二氧化硅	风化前	14	87.05	59.01	67.028	8.415	63.825	70.82	1.151	1.38	0.126
(SiO2)	风化后	6	96.77	92.35	93.963	1.734	93.505	3.005	-0.388	0.854	0.018
氧化钠	风化前	14	3.38	0	0.976	1.4	0	1.959	-1.215	0.859	1.433
(Na2O)	风化后	6	0	0	0	0	0	0	0	0	0.000
氧化钾	风化前	14	14.52	0	8.937	3.758	9.545	14.121	1.142	-0.848	0.420

根据 9 项统计指标观察统计数据，氧化钾 (K2O), 氧化钙 (CaO), 氧化镁 (MgO), 氧化铝 (Al2O3), 氧化铁 (Fe2O3), 氧化铜 (CuO), 氧化铅 (PbO), 五氧化二磷 (P2O5) 这些成分在风化过程中有存在部分的流失，但是在风化后还是会留有部分剩余，柱状图展示结果见图7。

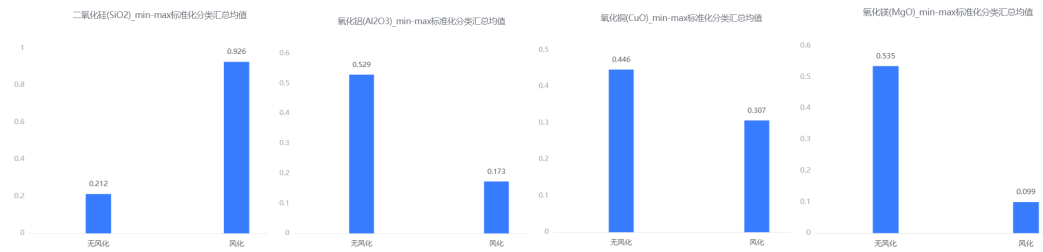


图 **7**　部分成分柱状图显示

接下来通过正态性检验对数据进行 Shapiro-Wilk（小数据样本，一般样本数 5000 以下）或者 Kolmogorov-Smirnov（大数据样本，一般样本数 5000 以上）检验，查看其显著性；分析结果中如若不呈现出显著性（p 值大于 0.05 或 0.01，严格为 0.05，不严格为 0.01），说明符合正态分布，反之说明不符合正态分布；

图8展示了经过标准化后的数据的正态性检验直方图，正态图基本上呈现出钟型（中间高，两端低），则说明数据虽然不是绝对正态，但是基本可以接受为正态分布。

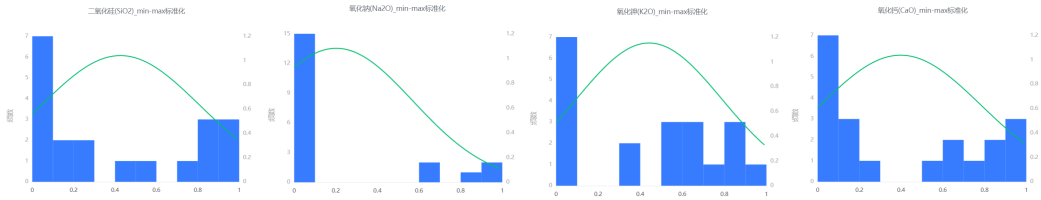


图 **8**　部分成分正态性检验直方图显示

5.2.3 建立多元线性回归模型

在数据预处理中，由于玻璃类型、风化、颜色等指标为定类变量，因此，需要通过独热编码对这些指标进行编码，使其转化为二进制。

多元线性回归 (multiple linear regression) 是研究一个连续型因变量和多个自变量之间线性关系的统计学分析方法。

1. 模型多元线性回归分析的模型为

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon, \\ \varepsilon \sim N\left(0, \sigma^2\right), \end{cases} \tag{1}$$

式中: $\beta_0, \beta_1, \cdots, \beta_m, \sigma^2$ 都是与 $x_1, x_2, \cdots, x_m$ 无关的未知参数, $\beta_0, \beta_1, \cdots, \beta_m$ 称为回归系数。

现得到 n 个独立观测数据 $[b_i, a_{i1}, \cdots, a_{i\,\mathrm{m}}]$, 其中 b_i 为 y 的观察值, $a_{i1}, \cdots, a_{im}$ 分别为 $x_1, x_2, \cdots, x_m$ 的观察值, $i = 1, \cdots, n, n > m$, 由式 (1) 得

$$\begin{cases} b_i = \beta_0 + \beta_1 a_{i1} + \cdots + \beta_m a_{im} + \varepsilon_i, \\ \varepsilon_i \sim N\left(0, \sigma^2\right), \quad i = 1, \cdots, n。 \end{cases} \tag{2}$$

记

$$\boldsymbol{X} = \begin{bmatrix} 1 & a_{11} & \cdots & a_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & a_{n1} & \cdots & a_{nm} \end{bmatrix}, \quad \boldsymbol{Y} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}, \tag{3}$$
$$\boldsymbol{\varepsilon} = [\varepsilon_1, \cdots, \varepsilon_n]^{\mathrm{T}}, \boldsymbol{\beta} = [\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m]^{\mathrm{T}},$$

式 (1) 表示为

$$\begin{cases} \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \\ \boldsymbol{\varepsilon} \sim N\left(0, \sigma^2 \boldsymbol{E}_n\right), \end{cases} \tag{4}$$

式中: $\boldsymbol{E}_n$ 为 n 阶单位矩阵。

3. 统计分析

不加证明地给出以下结果:

（1）$\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{\beta}$ 的线性无偏最小方差估计; $\hat{\boldsymbol{\beta}}$ 的期望等于 $\boldsymbol{\beta}$; 在 $\boldsymbol{\beta}$ 的线性无偏估计中, $\hat{\boldsymbol{\beta}}$ 的方差最小。

(2) $\hat{\boldsymbol{\beta}}$ 服从正态分布

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\right) \tag{5}$$

记 $\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1} = (c_{ij})_{n\times n}$。

通过代码求解，可以得出 F 检验的分析描述，例如：从 F 检验的结果分析可以得到，显著性 P 值为 0.000***，水平上呈现显著性, 拒绝回归系数为 0 的原假设，因此模型基本满足要求。

求得的结果详见附件 “风化文物未风化点风化前的化学成分含量.csv”

5.3 问题二的建模与求解

需要我们以高钾玻璃和铅钡玻璃两种类型进行进一步分类以及亚类划分, 并且针对具体分类模型作出合理性和敏感性分析，我们建立玻璃类型的整体聚类模型，在此基础上根据标准差进行二类划分，问题二建模流程如下图5所所示:

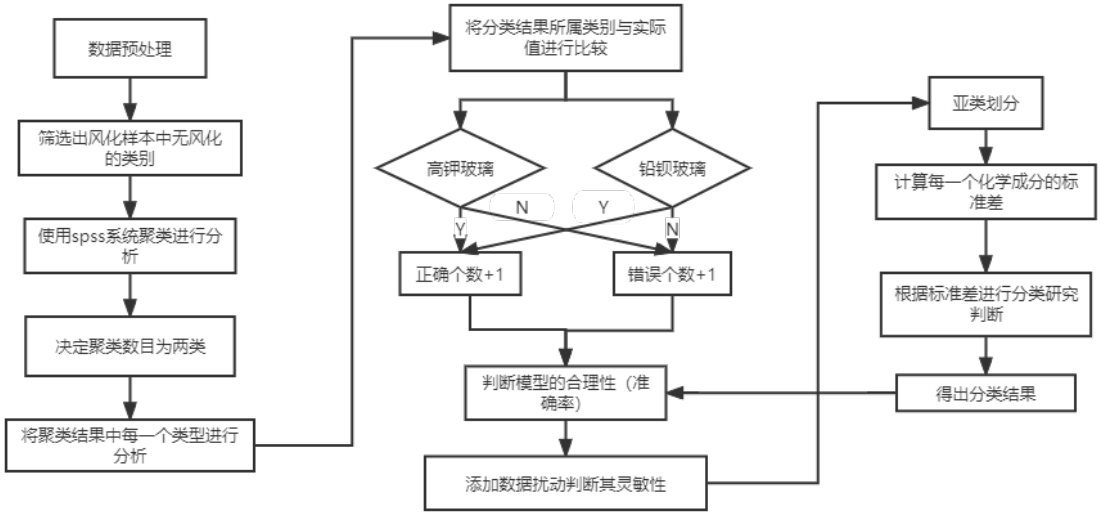


图 **9**　问题二求解流程图

5.3.1 数据预处理

首先，通过 Python 代码将附件中的相关数据进行提取，重新生成新文件。其中包含玻璃类型以及所有化学成分。通过查看文件可以看出文件只包含了高钾玻璃和铅钡玻璃两种。通过筛选，分别得出高钾玻璃数据集和铅钡玻璃数据集。

5.3.2 分析两种玻璃的分类规律

根据问题一求解得到的不同类型玻璃中风化前后的特征值，通过进行分类汇总，分析两种玻璃的分类规律。

类型	表面风化	二氧化硅(SiO2)	氧化钠(Na2O)	氧化钾(K2O)	氧化钙(CaO)	氧化镁(MgO)	氧化铝(Al2O3)	氧化铁(Fe2O3)	氧化铜(CuO)	氧化铅(PbO)	氧化钡(BaO)	五氧化二磷(P2O5)	氧化锶(SrO)	氧化锡(SnO2)	二氧化硫(SO2)
铅钡	无风化	14.587	1.538	0.398	1.458	0.545	1.385	1.445	2.491	9.094	6.950	1.571	0.314	0.158	1.015
	风化	17.221	1.920	0.215	1.612	0.656	3.413	0.692	2.488	15.164	8.868	4.151	0.245	0.232	3.609
高钾	无风化	8.415	1.400	3.758	3.441	0.625	2.998	1.567	1.595	0.548	0.929	1.387	0.047	0.631	0.175
	风化	1.734	0.000	0.445	0.488	0.306	0.964	0.069	0.935	0.000	0.000	0.210	0.000	0.000	0.000

图 **10**　两种玻璃的分类汇总

选取部分化学成分进行柱形图展示，显示效果如图10

通过显示的部分主要化学成分可以看出在铅钡玻璃更加容易风化。

5.4 聚类划分

5.4.1 聚类算法的概念

聚类算法：一种典型的无监督学习算法，主要用于将相似的样本自动归到一个类别中。

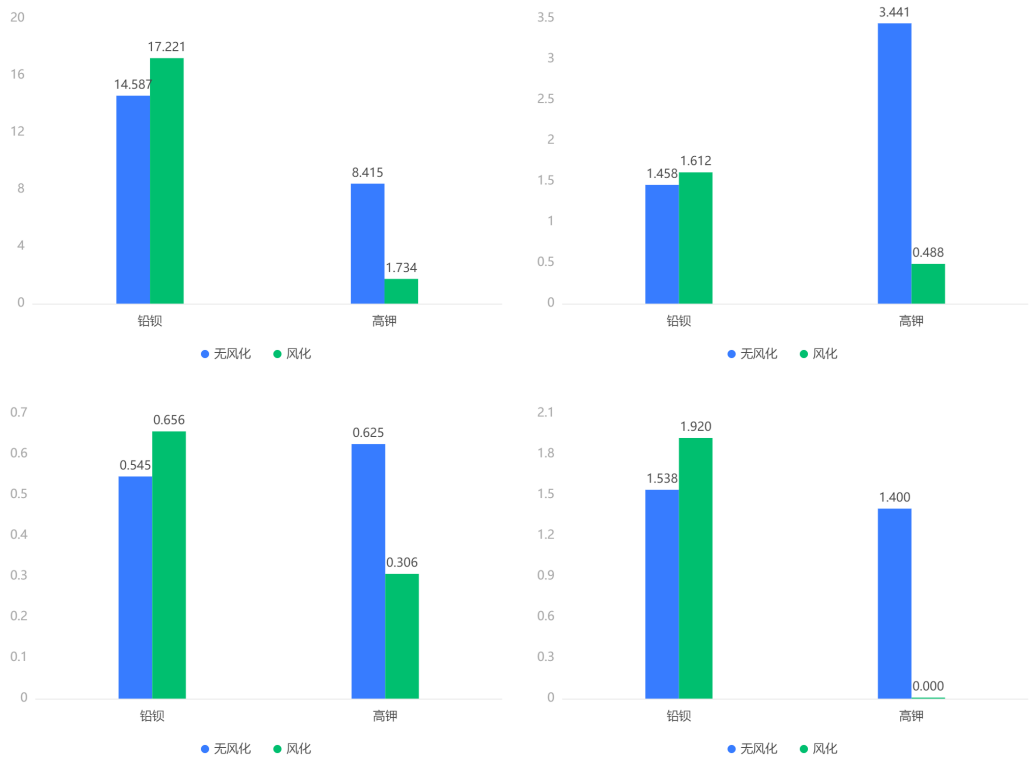


图 **11**　铅钡、高钾玻璃有无风化部分化学成分统计分析

在聚类算法中根据样本之间的相似性，将样本划分到不同的类别中，对于不同的相似度计算方法，会得到不同的聚类结果，常用的相似度计算方法有欧式距离法。

聚类算法是无监督的学习算法，而分类算法属于监督的学习算法。

5.4.2 k-means 聚类步骤

1. 随机设置 K 个特征空间内的点作为初始的聚类中心
2. 对于其他每个点计算到 K 个中心的距离，未知的点选择最近的一个聚类中心点作为标记类别
3. 接着对着标记的聚类中心之后，重新计算出每个聚类的新中心点（平均值）
4. 如果计算得出的新中心点与原中心点一样（质心不再移动），那么结束，否则重新进行第二步过程

5.4.3 “肘” 方法-K 值确定

$$\sum_{i=0}^{n} \min_{\mu_j \in C} \left(\|x_i - \mu_j\|^2 \right) \tag{6}$$

一般情况下会计算 K 值从 2-10 的情况，然后得出上述的 elbow 图，最后选择最优的那个 K 值。

（1）对于 n 个点的数据集，迭代计算 k from 1 to n，每次聚类完成后计算每个点到其所属的簇中心的距离的平方和；

（2）平方和是会逐渐变小的，直到 k==n 时平方和为 0，因为每个点都是它所在的簇中心本身。

（3）在这个平方和变化过程中，会出现一个拐点也即“肘”点，下降率突然变缓时即认为是最佳的 k 值。

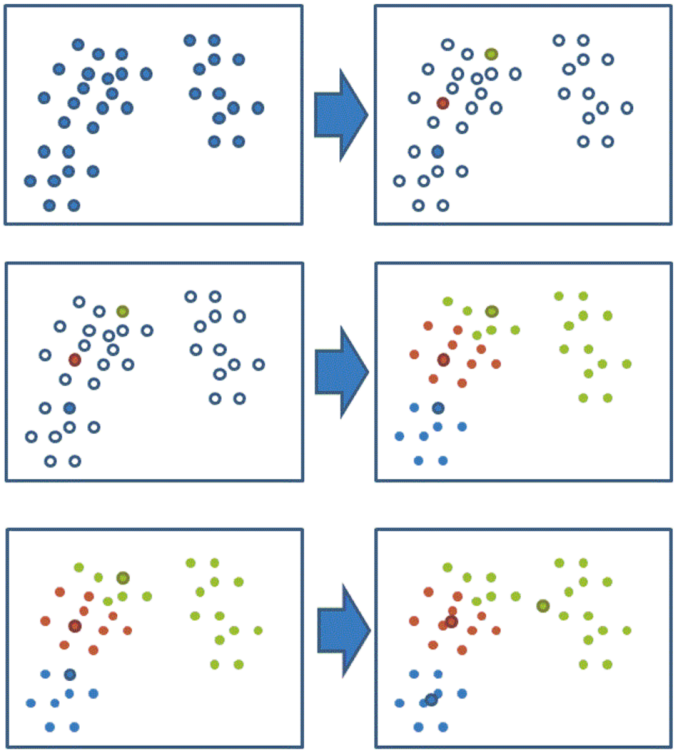


图 **12**　**K-means** 过程分析

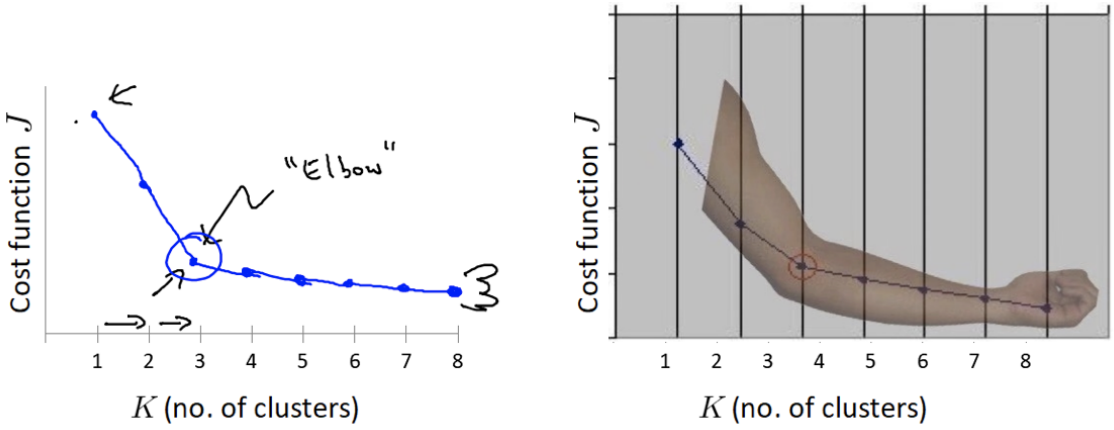


图 **13**　“肘”方法确定 **k** 值

在决定什么时候停止训练时，肘形判据同样有效，数据通常有更多的噪音，在增加分类无法带来更多回报时，我们停止增加类别。由图14可以看出聚类分析中，k值为 3 或者 4 时能够取得比较好的效果。

5.5 问题三的建模与求解

5.5.1 数据集特征生成

首先需要需要从附件表单二中提取出类型、表面风化以及化学成分等特征数据，然后需要对表面风化定类特征进行独热编码，从而生成两个新特征。

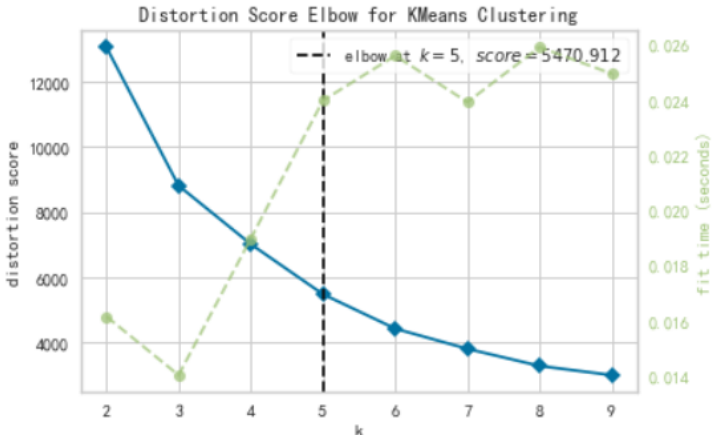


图 **14** **Kmeans** 聚类分析

5.5.2 模型建立

逻辑回归 (Logistic Regression）模型是一种对数线性模型，模型解释性强且原理易理解，它是统计学中的经典分类方法。

逻辑回归模型的分布函数和密度函数分别为:

$$F(x) = P(X \leq x) = \frac{1}{1 + exp(-(x-\mu)/\gamma)} \tag{7}$$

$$f(x) = F'(x) = \frac{exp(-(x-\mu)/\gamma)}{\gamma(1 + exp(-(x-\mu)/\gamma))^2} \tag{8}$$

其中，μ 为位置参数，$\gamma > 0$ 为形状参数。逻辑回归模型的分布函数是一条 S 型曲线，参数 γ 的值越小，曲线在中心处位置增长的就越快。

逻辑回归模型可以做二分类和多分类, 本文主要介绍逻辑回归的二分类算法。逻辑回归模型二分类算法的条件概率分布如下:

$$P(Y = 1|x) = \frac{exp(\theta x + b)}{1 + exp(\theta x + b)} \tag{9}$$

$$P(Y = 0|x) = \frac{1}{1 + exp(\theta x + b)} \tag{10}$$

其中，x 为特征变量，$Y \in 0, 1$ 为目标变量，θ 为权值向量，b 为偏置。由式 (6) 和式 (7) 可得

$$log\frac{P(Y = 1|x)}{1 - P(Y = 0|x)} = \theta x + b \tag{11}$$

逻辑回归模型训练时，对于给定的训练数据集 D，设 $P(Y = 1|x) = \varphi(x), P(Y = 1|x) = 1 - \varphi(x)$，似然函数为

$$\prod_{i=1}^{N} [\varphi(X_i)]^{y_i} [\varphi(X_i)]^{1-y_i} \tag{12}$$

对数似然函数为

$$L(\theta) = \sum_{i=1}^{N}[y_i log\varphi(x_i) + (1-y_i)log(1-\varphi(x_i))] \tag{13}$$

$$= \sum_{i=1}^{N}[y_i log\frac{\varphi(x_i)}{1-\varphi(x_i)} + log(1-\varphi(x_i))] \tag{14}$$

$$= \sum_{i=1}^{N}[y_i(\theta x_i) - log(1+e^{\theta x_i})] \tag{15}$$

对 $L(\theta)$ 求极大值，得到 θ 的估计值 $\hat{\theta}$。

$$h(w) = w_1x_1 + w_2x_2 + w_3x_3 \ldots + b \tag{16}$$

激活函数

- sigmoid 函数

$$g\left(w^T, x\right) = \frac{1}{1+e^{-h(w)}} = \frac{1}{1+e^{-w^Tx}} \tag{17}$$

- 判断标准
  - 回归的结果输入到 sigmoid 函数当中
  - 输出结果：[0, 1] 区间中的一个概率值，默认为 0.5 为阈值

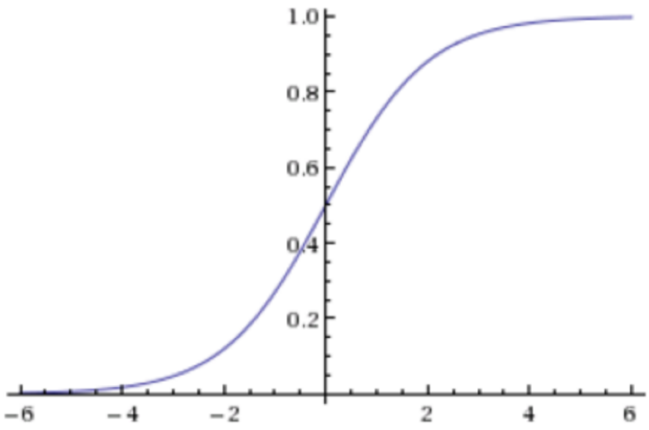


图 **15** **sigmoid** 图像

逻辑回归最终的分类是通过属于某个类别的概率值来判断是否属于某个类别，并且这个类别默认标记为 1(正例), 另外的一个类别会标记为 0(反例)。[3]

下图中，设置阈值为 0.6

5.5.3 数据集划分

经过数据集划分，将原始数据划分为训练数据和测试数据，建立模型之后，通过对训练集进行训练，得到基础模型，然后通过测试集进行模型预测。

5.5.4 模型评估

表4显示了对逻辑回归模型进行预测的结果，总体上模型效果优良，预测效果较好。

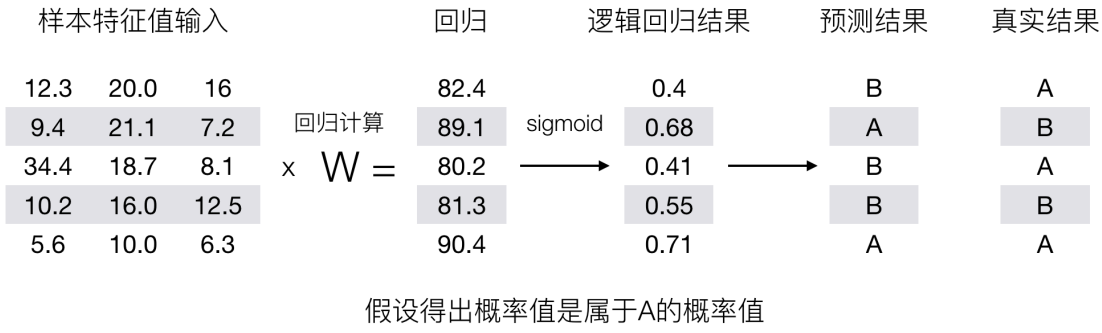


图 **16**　逻辑回归运算过程

表 **4**　逻辑回归模型分析

逻辑回归	precision	recall	precsion	F1-Socre
训练集	1	1	1	1
交叉验证集	0.75	0.75	0.78	0.747
测试集	0.905	0.905	0.905	0.905

5.5.5 预测模型评价指标

模型性能评价指标是选择模型优劣的重要参考，评价指标主要有准确率、混淆矩阵、F1 分数以及 ROC 曲线。

- 准确率

  评价分类器性能的指标一般是分类准确率 (accuracy)，是用来表示模型分类正确总度量。它的计算方式为测试集样本数据中模型预测结果预测正确数量之和比上测试集样本总量，即:

$$准确率 = \frac{\sum_{i=1}^{n} 预测结果类别为i且实际类别也为i的样本数量}{总样本量} \tag{18}$$

  精度能够直观的判断分类模型的分类准确程度, 但在判断之前需要确认样本数据是否存在样本过采样、欠采样问题。比如说数据集样本总量为 N，有 i 个类别，其中某一类别 I 占数据总量的 90%，如果模型预测结果类别全为 1，模型的预测的准确率则为 90%。

- F1-Score

  从计算公式可以看出精确率与召回率之间有着某种联系, 精确率高时召回率相对会低，因此延伸出精确率与召回率的调和均值 F1，即

$$\frac{2}{F1} = \frac{1}{P} + \frac{1}{R} F1 = 2\frac{R*P}{R+P} = \frac{2TP}{2TP+FP+FN} \tag{19}$$

  精确率 P 和召回率 R 他们的值都高时，F1 值也高。

- 交叉验证

  交叉验证是验证分类器性能的一种统计分析方法，它的工作原理是将原始数据集分为训练集和验证集，用训练集进行训练模型，然后使用验证集对模型进行验证，以此作为评价分类器性能的指标。

  Hold-Out Method 方法是将数据集随机分为训练集和验证集两组，一组用来训练模型，一组用来验证模型的性能，此方法因没有达到交叉验证的效果并不常用。

Double Cross Validation 思想是先将数据集分为大小相等的两个子集，一个子集作为训练集一个作为测试集训练分类器,然后将训练集与测试集对换在此训练模型，两次训练集的辨识度作为输出的结果，其中不足之处在于训练集样本的分布不足以代表数据集总体样本分布。[8]

K-fold Cross Validation 思想是将数据集分成 K 组，在 K 组中每个子集做一次验证集，余下的 K-1 组用来训练分类器，得到 K 个训练模型。

六、 模型检验

6.1 敏感性分析

在模型建立之后，需要对模型进行敏感性分析。敏感性分析（sensitivity analysis）是指从定量分析的角度研究有关因素发生某种变化对某一个或一组关键指标影响程度的一种不确定分析技术。

具体分析步骤如下：

- 确定指标敏感性分析的对象是具体的技术方案及其反映的经济效益。因此，技术方案的某些经济效益评价指标
- 计算该技术方案的目标值一般将在正常状态下的经济效益评价指标数值，作为目标值。

敏感性分析结果如图17。

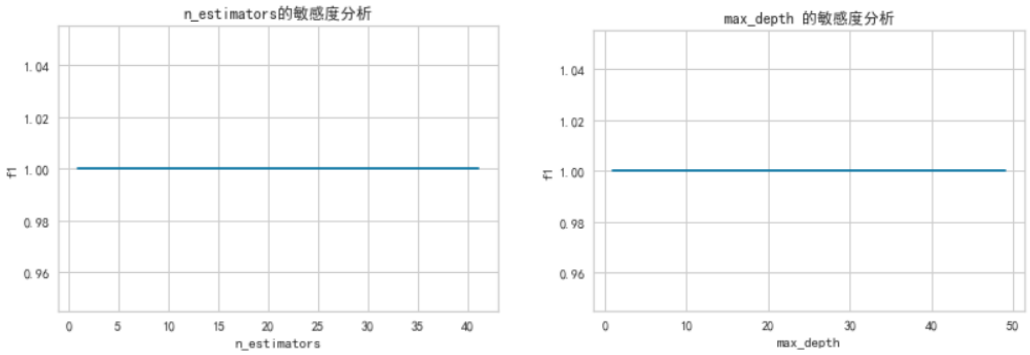


图 **17**　逻辑回归敏感性分析结果展示

6.2 差异性分析

图中展示模型检验的结果表分析了相关系数的正负向以及相关性程度进一步推出了其中的差异性若呈现数绝对值值较大，说明其两变量存在相关性，否则不相关

图18展示了热力图的形式展示了相关系数的值，主要通过颜色深浅去表示值的大小。

七、 模型评价与推广

7.1 Spearman 相关性系数分析

优点:

1. 既可以使用线性相关系数又适用于非线性相关系数
2. 在变量值没有变化的情况下，也不会出现像皮尔森系数那样分母为 0 而无法计算的情况。另外，即使出现异常值，由于异常值的秩次通常不会有明显的变化
3. 可以测量两个定序数据

图 **18**　相关系数

7.2 逻辑回归

优点：

1. 模型训练速度快，计算量仅仅只和特征的数目相关
2. 模型易于理解，解释性好，从特征的权重可以看到不同的特征对最后结果的影响
3. 适用于二分类问题，特征不需要缩放
4. 内存占用资源小

缺点：

1. 数据不平衡时不能得到处理
2. 准确率较低，形式简单以至难以去拟合数据的真实分布
3. 逻辑回归无法筛选其特征，gbdt 后才可以逻辑回归
4. 对多重共线性数据过为敏感

参考文献

[1] 刘露诗. 基于机器学习与缺失值插补技术的海底硫化物成矿定量预测 [D]. 吉林大学,2022.DOI:10.27162/d.cnki.gjlin.2022.000002.

[2] 卓炜杰. 函数型聚类分析方法及其应用研究 [D]. 浙江工商大学,2018.DOI:10.27462/d.cnki.ghzhc.2018.000045.

[3] 王凌妍, 张鑫雨, 许胜楠, 王禹力, 甄志龙. 逻辑回归的敏感性分析及在特征选择中的应用 [J]. 信息记录材料,2022,23(07):30-33.DOI:10.16009/j.cnki.cn13-1295/tq.2022.07.051.

[4] 于群, 霍筱东, 何剑, 李琳, 张建新, 冯煜尧. 基于斯皮尔曼相关系数和系统惯量的中国电网停电事故趋势预测 [J/OL]. 中国电机工程学报:1-12[2022-09-18].http://kns.cnki.net/kcms/detail/11.2107.TM.20220824.1625.012.html

[5] Collins M, Schapire R E, Singer Y.Logistic Regression, AdaBoost and Bregman Distances[J]. Machine Learning Journal,2002,48(1-3):253-285.

[6] 李航. 统计学习方法 [M]. 北京: 清华大学出版社,2012:116-123.

[7] Varshneya Arun K.,Macrelli Guglielmo,Yoshida Satoshi,Kim Seong H.,Ogrinc Andrew L.,Mauro John C.. Indentation and abrasion in glass products: Lessons learned and yet to be learned[J]. International Journal of Applied Glass Science,2022,13(3).

[8] 卢树强. 数学建模的算法创新与实践应用——评《数学建模算法与应用 (第 3 版)》[J]. 现代雷达,2022,44(03):111.

附录 A　材料结构

文件夹名	备注
源代码	附件及程序生成的数据
spsspro 分析结果	通过 SPSSPRO 软件生成的分析结果

附录 B　Python 源代码

```
import pandas as pd
import warnings
warnings.filterwarnings('ignore')

import os

data=pd.read_excel('附件.xlsx',sheet_name='表单1')
data.head()

data.columns

data.describe()

data.isnull().sum()

data.dropna().to_csv('Q1-1.csv',index=None,encoding='ANSI')

data2=pd.read_excel('附件.xlsx',sheet_name='表单2')

data2.head()

data2['文物编号']=data2['文物采样点'].apply(lambda x: int(str(x)[:2]))
data['颜色'].fillna(data['颜色'].mode()[0], inplace=True)
data2.isnull().sum()

data2.fillna(0, inplace=True)
data2.isnull().sum()
# 合并表格
data_merged=pd.merge(data,data2,on=['文物编号'])
data_merged.head()

data_merged.to_csv('Q1-2.csv',index=None,encoding='ANSI')
data_merged.columns

data_merged['成分总和']=0
for i in ['二氧化硅(SiO2)', '氧化钠(Na2O)',
'氧化钾(K2O)', '氧化钙(CaO)', '氧化镁(MgO)', '氧化铝(Al2O3)', '氧化铁(Fe2O3)',
'氧化铜(CuO)', '氧化铅(PbO)', '氧化钡(BaO)', '五氧化二磷(P2O5)', '氧化锶(SrO)',
'氧化锡(SnO2)', '二氧化硫(SO2)']:
data_merged['成分总和']+=data_merged[i]

data_merged[(data_merged['成分总和']>85)&data_merged['成分总和']<105]

data_merged.reset_index(inplace=True,drop=True)
data_merged.head()
```