古代玻璃制品的成分分析与鉴别模型

摘　要

本文通过数据分析的方法研究风化的玻璃文物和无风化的玻璃文物的表面特征与化学成分，从而给玻璃文物进行更加细致的划分得到不同的亚类，以及预测玻璃文物风化前后化学成分含量的变化，并对玻璃文物分出的不同亚类之间进行相关性与差异性的分析。

针对问题一，主要解决了三个问题：一是对玻璃文物的表面风化与其玻璃类型、纹饰和颜色进行相关性和差异性分析；二是统计出有无风化的文物样品表面的化学成分含量的规律；三是通过统计规律以及样本数据，预测出文物风化前的化学成分。鉴于表单一所给数据都是定类变量，所以我们选择了用于研究定类变量间关系的斯皮尔曼等级相关（Spearman rank correlation）和卡方检验，最终得出表面风化与玻璃类型相关性低，与纹饰和颜色具有显著相关性。在对表单 2 数据进行描述性统计时，通过对比风化前后各化学成分含量的平均值，中位数，方差，最大最小值，得出统计规律，即文物样品风化后，二氧化硅含量显著变多，氧化铅含量略微变多等等。根据统计规律以及表单 2 的数据预测出文物风化前的化学成分含量（见附件 1）。

针对问题二，主要解决了二个问题：一是找出了高钾玻璃和铅钡玻璃的分类规律；二是在高钾和铅钡两种类型的基础上继续分类，划分出亚类。分析问题可以发现，是否是高钾玻璃是二分类变量，典型的二元决策问题，所以我们采用了二元 logistic 回归，最终得到二氧化硅含量高，氧化钾含量高，氧化锶含量低玻璃文物，大概率是高钾类型的玻璃文物；反之则为铅钡类型的玻璃文物。对于数据量多的分类问题，采用 K-means 聚类算法进行亚类分类，聚类分析结果将高钾玻璃分为两个亚类，铅钡玻璃也是分为两个亚类。

针对问题三，在问题二得出的分类规律的基础上，先采用分类规律对未知类别的玻璃文物分类，得出属于高钾类型亦或者铅钡类型，再将具体的化学成分数据放入 K-means 聚类模型中，对表单 3 的玻璃文物进行更细致的亚类划分。得出结论 A1 属于高钾二类，A2 属于铅钡二类（具体见正文分析）。

针对问题四，主要解决了二个问题：一是分析了各类型的玻璃文物的化学成分之间的关联关系；二是对不同类别之间的化学成分关联关系的差异性分析。对于关联关系，我们选择灰色关联分析模型，选择其中一种作为参考量与其它变量进行灰色关联分析，得出参考量与其它变量的灰色关联系数与灰色关联度。差异性分析我们采用配对样本 T 检验，得到高钾玻璃和铅钡玻璃的二氧化硅和氧化铝的关联关系差异性幅度较小，不存在显著性差异。

本文采用了多种机器学习的方法，深入研究了玻璃文物的化学成分，表面基本特征对文物分类的相关性。

关键词：斯皮尔曼等级相关　卡方检验　K-means 聚类　二元 logistic 回归　灰色关联分析

一、 问题重述

古代玻璃受环境的影响产生风化，其内化学成分比例也从而改变。现需将一批高钾玻璃和铅钡玻璃的数据进行分析建模，解决以下问题。

问题 1：分析玻璃风化与玻璃表面特征的关系；通过风化与化学成分含量之间的关系，预测风化前的化学成分含量。

问题 2：高钾玻璃与铅钡玻璃的显著差别；在单种玻璃下再进行划分，并分析划分的合理性和敏感性。

问题 3：通过化学成分判断玻璃类型，并分析敏感性。

问题 4：分析不同类别化学成分之间的关联，和不同关联的差异性。

二、 问题分析

2.1 问题一的分析

先对数据进行处理，排除成分不在 85%~105%的无效数据。

风化与玻璃表面特征的关系：根据附件表单 1 的数据，使用 Spearman 相关性分析和卡方检验对玻璃不同特征进行相关性和差异性分析。

化学成分含量的统计规律：将两种玻璃和有无风化分成四组数据，再针对其中不同化学成分进行描述性统计分析，分析不同类型风化前后含量比例其有无、增减的变化。

预测风化前的化学成分含量：根据风化前后各个化学成分的含量比例变化去推导风化前的含量。

2.2 问题二的分析

高钾和铅钡玻璃的分类：分析两种玻璃的化学成分含量风化前后的比例和比例变化的差异，结合逻辑回归分类。

在各个类别再选择化学成分进行亚类划分：在不同类别玻璃内，选择对该种玻璃类别判断影响较大的化学成分进行 KMeans 聚类算法进行亚类划分。

2.3 问题三的分析

在问题二里我们得出了高钾玻璃与铅钡玻璃的分类规律，那么我们可以通过观察表单 3 中文物的化学成分来得出其属于高钾玻璃还是铅钡玻璃。得到文物的类型后，我们再根据 K-means 聚类算法将其进行亚类划分，得出文物所属亚类。

2.4 问题四的分析

分析关联关系：我们采用灰色关联分析模型来分析高钾玻璃和铅钡玻璃的化学成分的关联关系，在问题一里我们可以知道对高钾玻璃和铅钡玻璃有较大影响的化学成分，然后选择其中一种作为参考量与其它变量进行灰色关联分析，得出参考量与其它变量的灰色关联系数与灰色关联度。

比较不同类别之间的化学成分关联关系的差异性：采用配对样本 T 检验，对得到的关联系数进行差异性分析。

三、 模型假设

1. 附件里的 3 个表单的数据真实可信

2. 假设附件所给数据量已经足够，无需再增设新数据
3. 假设玻璃文物的化学成分含量只会受到风化的影响，不考虑其他因素对化学成分的影响

四、 符号说明

符号	说明
ρ	斯皮尔曼等级相关系数
Y	玻璃文物表面是否风化的集合
X	玻璃类型、纹饰和颜色的集合
x	X 排序后的新集合
y	Y 排序后的新集合
d	x、y 中的元素对应相减得到排行差分集合
χ^2	卡方检验的统计值
$O_{i,j}$	卡方检验的理论值
$E_{i,j}$	卡方检验的观察值
x_i	玻璃类型的化学成分的第 i 个解释变量
k	解释变量个数
α	截距项
β_i	x_i 的系数
A	包含了 n 个点的簇类
a_n	表示 A 的第 n 个对象
$dis(A,B)$	某个对象到聚类中心的欧氏距离

五、 模型的建立与求解

5.1 问题一模型的建立与求解

5.1.1 关系分析

对玻璃文物的表面分化与其玻璃类型、纹饰和颜色的关系进行相关性分析，使用 Spearman rank correlation 对所给玻璃文物特征数据进行相关性分析。

（1）Spearman 相关性分析

在统计学中，[1]斯皮尔曼等级相关（Spearman rank correlation）是用来估算两个变量之间的相关性，它要求两个变量的观测值是成对的等级评价资料，或由连续变量观测资料转化得到的等级资料，其优点是无需考虑两个变量之间的总体分布形态和样本容量的多少。对于本题，将玻璃文物表面是否风化作为变量 Y，将其表面特征数据，即玻璃类型、纹饰和颜色逐一作为变量 X，去掉样本数据中的 4 个有缺失值的样本，样本数据的元素个数均为 n（n=54），两个变量取得第 i($1\leq i \leq n$)个值分别表示为 X_i、Y_i，对 X、Y 同时进行升或降排序，分别得到两个对元素进行排序后的新集合 x、y，其中元素 x_i 为 X_i 在 X 中的排行、y_i 为 Y_i 在 Y 中的排行.将集合 x、y 中的

元素对应相减得到一个排行差分集合 d，其中 $d_i=x_i-y_i, 1\leq i \leq n$。则随机变量 X、Y 之间的斯皮尔曼等级相关系数定义为

$$\rho = 1 - \frac{6\sum_i^{\infty} d_i^2}{n(n^2-1)} \quad (1)$$

通过公式（1），将样本数据代入，我们可以得到表面风化与玻璃类型、纹饰、颜色的斯皮尔曼系数。

表 1 表面风化与类型

			类型	表面风化
斯皮尔曼 Rho	类型	相关系数	1.000	-.316*
		显著性（双尾）	.	.020
		个案数	54	54
	表面风化	相关系数	-.316*	1.000
		显著性（双尾）	.020	.
		个案数	54	54

*. 在 0.05 级别（双尾），相关性显著。

根据观察表 1，它们之间的斯皮尔曼系数显示为-0.316，说明玻璃类型对表面风化存在一定的负向相关性。

表 2 表面风化与纹饰

			表面风化	纹饰
斯皮尔曼 Rho	表面风化	相关系数	1.000	.048
		显著性（双尾）	.	.731
		个案数	54	54
	纹饰	相关系数	.048	1.000
		显著性（双尾）	.731	.
		个案数	54	54

根据观察表 2，它们之间的斯皮尔曼系数显示为 0.048，说明玻璃纹饰对表面风化存在一定的正向相关性。

表 3 表面分化与颜色

			表面风化	颜色
斯皮尔曼 Rho	表面风化	相关系数	1.000	-.162
		显著性（双尾）	.	.243
		个案数	54	54
	颜色	相关系数	-.162	1.000
		显著性（双尾）	.243	.
		个案数	54	54

根据观察表 2，它们之间的斯皮尔曼系数显示为-0.162，说明玻璃颜色对表面风化存在一定的负向相关性。

通过使用斯皮尔曼等级相关对玻璃文物表面风化与表面特征进行相关性分析，但还不能确定它们的关系，需要再对它们的差异性进行分析。

（2）Pearson 卡方检验

[2]Pearson 卡方检验是最有名的卡方检验之一，主要是比较定类变量与定类变量之间的差异性。使用卡方检验对表单一的数据进行差异性分析，需要计算卡方检验

的统计值χ^2：把每个理论值和观察值的做差，然后平方、再除以理论值、最后求和：

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}} \tag{2}$$

在公式（2）中$O_{i,j}$是理论值，$E_{i,j}$是观察值。

通过公式（2），将样本数据代入，我们可以得到表面风化与玻璃类型、纹饰、颜色的卡方值。

表 4 表面风化与类型卡方检验

	值	自由度	渐进显著性（双侧）
皮尔逊卡方	5.400[a]	1	.020
连续性修正[b]	4.134	1	.042
似然比(L)	5.448	1	.020
线性关联	5.300	1	.021
有效个案数	54		

根据观察表 4，Pearson 卡方检验分析的结果显示，基于表面风化和类型，显著性 P 值为 0.020，水平上呈现显著性，拒绝原假设，所以对于类型和表面风化数据存在显著性差异。

表 5 表面风化与纹饰卡方检验

	值	自由度	渐进显著性（双侧）
皮尔逊卡方	5.747[a]	2	.056
似然比(L)	7.993	2	.018
线性关联	.205	1	.650
有效个案数	54		

a．2 个单元格（33.3%）的期望计数小于 5。最小期望计数为 2.67。

根据观察表 5，Pearson 卡方检验分析的结果显示，基于表面风化和纹饰，显著性 P 值为 0.056，水平上不呈现显著性，接受原假设，所以对于纹饰和表面风化数据不存在显著性差异。

表 6 表面风化与颜色卡方检验

	值	自由度	渐进显著性（双侧）
皮尔逊卡方	6.287[a]	7	.507
似然比(L)	8.156	7	.319
线性关联	2.616	1	.106
有效个案数	54		

a．12 个单元格（75.0%）的期望计数小于 5。最小期望计数为 .44。

根据观察表 6，Pearson 卡方检验分析的结果显示，基于表面风化和颜色，显著性 P 值为 0.507，水平上不呈现显著性，接受原假设，所以对于颜色和表面风化数据不存在显著性差异。

5.1.2 统计规律

对高钾玻璃和铅钡玻璃的化学成分含量比例进行描述性统计。

（1） 高钾玻璃描述性统计

表 7 高钾玻璃（未风化）统计数据

高钾（未风化）	样本量	平均值	标准差	中位数	方差	变异系数（CV）
二氧化硅(SiO2)	13	66.285	8.268	62.47	68.361	0.125
氧化钠(Na2O)	13	1.052	1.427	0	2.036	1.357
氧化钾(K2O)	13	9.625	2.851	9.67	8.13	0.296
氧化钙(CaO)	13	4.56	3.581	5.87	12.827	0.785
氧化镁(MgO)	13	1.046	0.649	1.02	0.421	0.62
氧化铝(Al2O3)	13	5.877	3.12	6.16	9.732	0.531
氧化铁(Fe2O3)	13	1.761	1.622	1.74	2.631	0.921
氧化铜(CuO)	13	2.195	1.632	2.18	2.664	0.744
氧化铅(PbO)	13	0.332	0.539	0.19	0.291	1.622
氧化钡(BaO)	13	0.401	0.863	0	0.745	2.153
五氧化二磷(P2O5)	13	1.244	1.443	0.79	2.083	1.16
氧化锶(SrO)	13	0.038	0.048	0	0.002	1.242
氧化锡(SnO2)	13	0.182	0.655	0	0.428	3.606
二氧化硫(SO2)	13	0.094	0.18	0	0.032	1.916

表 8 高钾玻璃（风化）统计数据

高钾（风化）	样本量	平均值	标准差	中位数	方差	变异系数（CV）
二氧化硅(SiO2)	6	93.963	1.734	93.505	3.005	0.018
氧化钠(Na2O)	6	0	0	0	0	0
氧化钾(K2O)	6	0.543	0.445	0.665	0.198	0.819
氧化钙(CaO)	6	0.87	0.488	0.83	0.238	0.561
氧化镁(MgO)	6	0.197	0.306	0	0.094	1.558
氧化铝(Al2O3)	6	1.93	0.964	1.72	0.93	0.5
氧化铁(Fe2O3)	6	0.265	0.069	0.275	0.005	0.262
氧化铜(CuO)	6	1.562	0.935	1.545	0.874	0.599
氧化铅(PbO)	6	0	0	0	0	0
氧化钡(BaO)	6	0	0	0	0	0
五氧化二磷(P2O5)	6	0.28	0.21	0.28	0.044	0.75
氧化锶(SrO)	6	0	0	0	0	0
氧化锡(SnO2)	6	0	0	0	0	0
二氧化硫(SO2)	6	0	0	0	0	0

①从表中可知，高钾玻璃在风化过程中，除二氧化硅的含量比例平均值增加外，其余均减少。

②未风化高钾玻璃氧化钠的变异系数较大，不适宜用平均数，从具体数据可知未风化 13 个样品中有 5 个含氧化钠，风化 6 个样品中均不含氧化钠，无法判断。

③未风化高钾玻璃氧化铅的变异系数较大，不适宜用平均数，从具体数据可知未风化 13 个样品中有 8 个含氧化铅，风化 6 个样品均不含氧化铅，猜测风化会使高钾玻璃的氧化铅含量比例减少。

④未风化高钾玻璃氧化钡的变异系数较大，不适宜用平均数，从具体数据可知未风化 13 个样品中有 3 个含氧化钡，风化 6 个样品中均不含氧化钡，猜测风化不会影响高钾玻璃的氧化钡含量比例。

⑤未风化高钾玻璃五氧化二磷的变异系数较大，不适宜用平均数，从具体数据可知未风化 13 个样品中有 1 个样品的五氧化二磷相较于其他样品高出许多，但其余样品的五氧化二磷含量比例相较于风化后的含量比例较高，推测风化会使高钾玻璃的五氧化二磷含量比例减少。

⑥未风化高钾玻璃氧化锶的变异系数较大，不适宜用平均数，从具体数据可知未风化 13 个样品中有 6 个含氧化锶，但含量比例都很小，风化 6 个样品中均不含氧化锶，猜测风化不会影响高钾玻璃的氧化锶含量比例。

⑦未风化高钾玻璃氧化锡的变异系数较大，不适宜用平均数，从具体数据可知未风化 13 个样品中仅有 1 个含氧化锡，风化 6 个样品中均不含氧化锡，猜测风化不影响高钾玻璃氧化锡的含量比例。

⑧未风化高钾玻璃二氧化硫的变异系数较大，不适宜用平均数，从具体数据可知未风化 13 个样品中有 3 个含二氧化硫，风化 6 个样品中均不含二氧化硫，猜测风化不会影响高钾玻璃的二氧化硫含量比例。

⑨风化高钾玻璃氧化镁的变异系数较大，不适宜用平均数，从具体数据可知未风化 13 个样品中有 11 个含氧化镁，变异系数 0.62，可参考平均值为 1.046；风化 6 个样品中有 2 个含氧化镁，含量比例分别为 0.64、0.54，其余均为 0。判断风化会使高钾玻璃的氧化镁含量比例减少。

（2）铅钡玻璃描述性统计

表 9 铅钡玻璃（未风化）统计数据

铅钡（未风化）	样本量	平均值	标准差	中位数	方差	变异系数（CV）
二氧化硅(SiO2)	13	53.444	14.587	55.21	212.788	0.273
氧化钠(Na2O)	13	0.772	1.538	0	2.366	1.994
氧化钾(K2O)	13	0.258	0.398	0.15	0.158	1.54
氧化钙(CaO)	13	1.232	1.458	0.84	2.125	1.184
氧化镁(MgO)	13	0.492	0.545	0.51	0.298	1.108
氧化铝(Al2O3)	13	3.195	1.385	3.06	1.919	0.434
氧化铁(Fe2O3)	13	0.933	1.445	0	2.088	1.549
氧化铜(CuO)	13	1.557	2.491	0.53	6.205	1.6
氧化铅(PbO)	13	23.594	9.094	22.05	82.708	0.385
氧化钡(BaO)	13	10.499	6.95	10.06	48.299	0.662
五氧化二磷(P2O5)	13	0.904	1.571	0.2	2.467	1.738

氧化锶(SrO)	13	0.297	0.314	0.3	0.098	1.057
氧化锡(SnO2)	13	0.065	0.158	0	0.025	2.444
二氧化硫(SO2)	13	0.282	1.015	0	1.03	3.606

表 10 铅钡玻璃（风化）统计数据

铅钡（风化）	样本量	平均值	标准差	中位数	方差	变异系数（CV）
二氧化硅(SiO2)	34	33.653	17.716	28.97	313.858	0.526
氧化钠(Na2O)	34	1.009	1.963	0	3.852	1.946
氧化钾(K2O)	34	0.12	0.151	0	0.023	1.256
氧化钙(CaO)	34	2.329	1.656	2.11	2.744	0.711
氧化镁(MgO)	34	0.69	0.67	0.71	0.449	0.971
氧化铝(Al2O3)	34	3.791	3.498	2.63	12.239	0.923
氧化铁(Fe2O3)	34	0.495	0.659	0.26	0.434	1.332
氧化铜(CuO)	34	2.003	2.531	1	6.404	1.264
氧化铅(PbO)	34	36.386	15.465	36.765	239.181	0.425
氧化钡(BaO)	34	10.947	8.891	8.79	79.049	0.812
五氧化二磷(P2O5)	34	4.035	4.193	2.835	17.584	1.039
氧化锶(SrO)	34	0.377	0.249	0.36	0.062	0.66
氧化锡(SnO2)	34	0.059	0.238	0	0.057	4.028
二氧化硫(SO2)	34	1.045	3.708	0	13.748	3.549

表中，红色为铅钡玻璃化学成分含量比例风化后平均值增加，绿色为铅钡玻璃化学成分含量比例风化后平均值减少。由于许多化学成分含量比例的变异系数都较大，所以采用平均数和中位数，和具体数据进行分析。

①二氧化硅的变异系数较少，从平均值和中位数可知，风化会使铅钡玻璃二氧化硅的含量比例减少。

②氧化钠的风化前后中位数均为 0，风化前后平均值相差 0.2，相对较小，且大多数数据的含量比例为 0，也有特别大的含量比例，猜测风化不影响氧化钠的含量比例。

③氧化钾的风化前后中位数均为 0，风化前后平均值相差 0.1，相对较小，且大多数数据的含量比例相对接近，猜测风化不影响氧化钾的含量比例。

④氧化钙的风化前后中位数和平均值相差均达到 1，且具体数据中含量比例为 0 的较少，大多数数据风化后较多，猜测风化会使铅钡玻璃氧化钙的含量比例增大。

⑤氧化镁的风化前后中位数和平均值相差 0.1，相对较小，且大多数数据的含量比例相对接近，猜测风化不影响氧化镁的含量比例。

⑥氧化铝的风化后比风化前中位数均减少 0.5，风化后比风化前平均值增高 0.5，具体数据有特别大的也有相近的，同时有一个样品风化点与未风化点氧化铝含量比例接近，猜测风化不影响氧化铝的含量比例。

⑦氧化铁具体数据中，风化前有一个数据较大，其余数据接近，猜测风化不影响氧化铁的含量比例。

⑧氧化铜风化后的中位数和平均值都有所增加，再根据具体数据风化后较风化前的数据都较大，猜测风化会使氧化铜的含量比例增大。

⑨氧化铅的变异系数较小，可根据平均数判断，猜测风化会使氧化铅的含量比例增大。

⑩氧化钡的变异系数较小，可根据平均数判断，猜测风化会使氧化钡的含量比例增大。

⑪五氧化二磷风化后的中位数与平均值都比风化前打，并根据具体数据分析可知，大多数数据风化后较风化前更大，猜测风化会使五氧化二磷的含量比例增大。

⑫氧化锶的风化前后中位数相差 0.05，风化前后平均值相差 0.05，相对较小，且大多数数据的含量比例接近，猜测风化不影响氧化锶的含量比例。

⑬氧化锡的风化前后中位数均为 0，风化前后平均值相差 0.01，相对较小，且大多数数据的含量比例为 0，猜测风化不影响氧化锡的含量比例。

⑭二氧化硫的风化前后中位数均为 0，风化前后平均值相差 0.8，但大多数数据的含量比例为 0，猜测风化不影响二氧化硫的含量比例。

5.1.3 预测风化前含量

依据统计规律推测风化前的化学成分含量比例。（预测结果见支撑材料/结论数据/Result1.xlsx）

表 11 高钾玻璃风化前预测公式

高钾玻璃	
二氧化硅	$Y_{SiO2}=X_{SiO2}-27.7$
氧化钾	$Y_{K2O}=X_{K2O}+9.1$
氧化钙	$Y_{CaO}=X_{CaO}+3.7$
氧化镁	$Y_{MgO}=X_{MgO}+0.8$
氧化铝	$Y_{Al2O3}=X_{Al2O3}+4$
氧化铁	$Y_{Fe2O3}=X_{Fe2O3}+1.5$
氧化铜	$Y_{CuO}=X_{CuO}+0.6$
氧化铅	$Y_{PbO}=X_{PbO}+0.3$
五氧化二磷	$Y_{P2O5}=X_{P2O5}+0.9$

表 12 铅钡玻璃风化前预测公式

铅钡玻璃	
二氧化硅	$Y_{SiO2}=X_{SiO2}+19.7$
氧化钙	$Y_{CaO}=X_{CaO}-1.1$
氧化铜	$Y_{CuO}=X_{CuO}-0.5$
氧化铅	$Y_{PbO}=X_{PbO}-12.8$

氧化钡	$Y_{BaO}=X_{BaO}-0.4$
五氧化二磷	$Y_{P2O5}=X_{P2O5}-3.1$

5.2 问题二模型的建立与求解

5.2.1 二元 logistic 回归

Logistic Regression 虽然被称为回归，但其实际上是分类模型，并常用于二分类。[3]为研究高钾玻璃、铅钡玻璃的分类规律，其中“类型里的高钾玻璃和铅钡玻璃”为被解释变量，且最终结果只有高钾玻璃或者铅钡玻璃两种，相当于“是”和“否”两个端点，属[0,1]二分类变量，是隶属于二元决策问题的典例，故我们可以采用二元 logistic 回归模型进行分析影响玻璃类型是高钾亦或者是铅钡的化学成分因素。如果是高钾玻璃则定义“y=0”，如果是铅钡玻璃则定义“y=1”。二元 logistic 模型的基本形式如：

$$\ln\left(\frac{p(y=1)}{1-p(y=1)}\right)=\alpha+\sum_{i=1}^{k}\beta i\, xi \tag{3}$$

模型中，x_i 表示影响玻璃类型的化学成分的第 i 个解释变量，k 为解释变量个数，α 为截距项，β_i 为 x_i 的系数，反映该变量对玻璃类型方向及程度，通常用最大似然估计法求得。玻璃类型是高钾玻璃的概率与玻璃类型是铅钡玻璃的概率的比值为事件发生比 p(y=1) /[1-p(y=1)]。

根据公式（3），选用玻璃类型为因变量，有无风化为分类协变量，逐一选取各个化学成分做为协变量，通过二元逻辑回归分析出哪些化学成分做为协变量时显著性高，即对玻璃类型具有一定相关性，分析结果如下：

表 13 二氧化硅逻辑回归

		B	标准误差	瓦尔德	自由度	显著性	Exp(B)
步骤 1[a]	二氧化硅(SiO2)	.169	.052	10.757	1	.001	1.184
	有无风化(1)	2.962	1.471	4.056	1	.044	19.345
	常量	-13.296	4.178	10.126	1	.001	.000

根据观察表 13，二氧化硅此协变量的显著性低于 0.005，说明二氧化硅与玻璃类型在统计学上具有统计显著，即二氧化硅对玻璃类型具有高度相关性，具此我们可猜测二氧化硅含量可能会影响玻璃类型。

表 14 氧化钾逻辑回归

		B	标准误差	瓦尔德	自由度	显著性	Exp(B)
步骤 1[a]	有无风化(1)	-1.149	1.272	.815	1	.367	.317
	氧化钾(K2O)	2.388	1.092	4.783	1	.029	10.887
	常量	-2.439	.591	17.043	1	.000	.087

根据观察表 14，氧化钾此协变量的显著性低于 0.05，说明氧化钾与玻璃类型在统计学上具有统计显著，即氧化钾对玻璃类型具有高度相关性，具此我们可猜测氧化钾含量可能会影响玻璃类型。

表 15 氧化锶逻辑回归

		B	标准误差	瓦尔德	自由度	显著性	Exp(B)
步骤 1[a]	有无风化(1)	1.281	.761	2.836	1	.092	3.601

氧化锶(SrO)	-13.906	4.394	10.017	1	.002	.000
常量	.100	.596	.028	1	.867	1.105

根据观察表 15，氧化锶此协变量的显著性低于 0.005，说明氧化锶与玻璃类型在统计学上具有统计显著，即氧化锶对玻璃类型具有高度相关性，具此我们可猜测氧化锶含量可能会影响玻璃类型。

在这里我们选择将 p 值小于 0.05 的表格呈现于论文上，即只呈现对玻璃类型有相关性的化学成分，其他化学成分的逻辑回归将置于附件。

表 16 描述性统计

	样本量	平均值	标准差	中位数	方差	变异系数（CV）
高钾（风化）						
二氧化硅(SiO2)	6	93.963	1.734	93.505	3.005	0.018
氧化钾(K2O)	6	0.543	0.445	0.665	0.198	0.819
氧化锶(SrO)	6	0	0	0	0	0
高钾（无风化）						
二氧化硅(SiO2)	13	66.285	8.268	62.47	68.361	0.125
氧化钾(K2O)	13	9.625	2.851	9.67	8.13	0.296
氧化锶(SrO)	13	0.038	0.048	0	0.002	1.242
铅钡（风化）						
二氧化硅(SiO2)	34	33.653	17.716	28.97	313.858	0.526
氧化钾(K2O)	34	0.12	0.151	0	0.023	1.256
氧化锶(SrO)	34	0.377	0.249	0.36	0.062	0.66
铅钡（无风化）						
二氧化硅(SiO2)	13	53.444	14.587	55.21	212.788	0.273
氧化钾(K2O)	13	0.258	0.398	0.15	0.158	1.54
氧化锶(SrO)	13	0.297	0.314	0.3	0.098	1.057

根据表 13、14、15，可以知道玻璃类型和 SiO2、K2O、SrO 这三个化学成分显著相关性，再结合表 16 对化学成分的描述性统计，可以知道二氧化硅含量高，氧化钾含量高，氧化锶含量低玻璃文物，大概率是高钾类型的玻璃文物；反之二氧化硅含量低，氧化钾含量低，氧化锶含量高的玻璃文物，大概率是铅钡类型的玻璃文物。

5.2.2 基于 K-means 聚类算法进行亚类分类

[4]K-means 聚类算法也称 k 均值聚类算法，是一种迭代求解的聚类分析算法，步骤是随机选取 k 个对象作为初始的聚类中心，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小。

K-means 聚类算法使用的数学公式如下：

$$A = (a_1, a_2, \dots, a_n), B = (b_1, b_2, \dots, b_n) \tag{4}$$

通过上面的公式我们可以得到 A 与 B 之间的距离公式：

$$dis(A,B) = \sqrt{\sum_{i=1}^{n}(b_i - a_i)^2} \tag{5}$$

将其化成矩阵的形式，可以得到如下公式：

$$(A-B)^T * (A-B) \tag{6}$$

具体的算法步骤如下：

① 从数据中随机挑选 k 个初始的样本中心点向量，记为 d1,d2,....,dk,，其中每个

向量都是 n 维的，和数据的特征维度相同；

② 计算每一个数据分别到这 k 个向量的距离，并把它归到最近的那个中心点上。这样我们就得到了每一个中心点的周围的数据点集合，将其归为一类，记为 Di；

③ 下面我们根据 Di 求出他们的质心，即他们真正的中心，替换了之前的样本中心点向量，这样就得到了新一轮的中心点，仍旧记为：d1,d2,....,dk；

④ 不断重复第二步和第三步，直到满足特定条件：例如说收敛，或是最大迭代轮次。直到获得最终版的 d1,d2,....,dk，；

⑤ 用 d1,d2,....,dk，去分堆，便可以得到 k 堆的 Di。

我们分别对高钾玻璃和铅钡玻璃进行亚类分类。

对于高钾玻璃，我们将附件表单 2 里所有属于高钾玻璃的化学成分的数据导入到高钾.xlsx 表里。

我们采用手肘法得出 k 值，手肘法将簇间误差平方和看成是类簇数量 k 的函数。随着 k 的增加，每个类簇内的离散程度越小，总距离平方和也就在不断减小，并且减小的程度越来越不明显。极限情况是当 k=N 时，每个类簇只有一个点，这时总的误差平方和为 0。手肘法认为我们应该选择这样的 k：当 k 继续增大时，总误差平方和减少的趋势不再明显，也就是“拐点”处。如图 1 所示，可以观察出 k 值为 2。

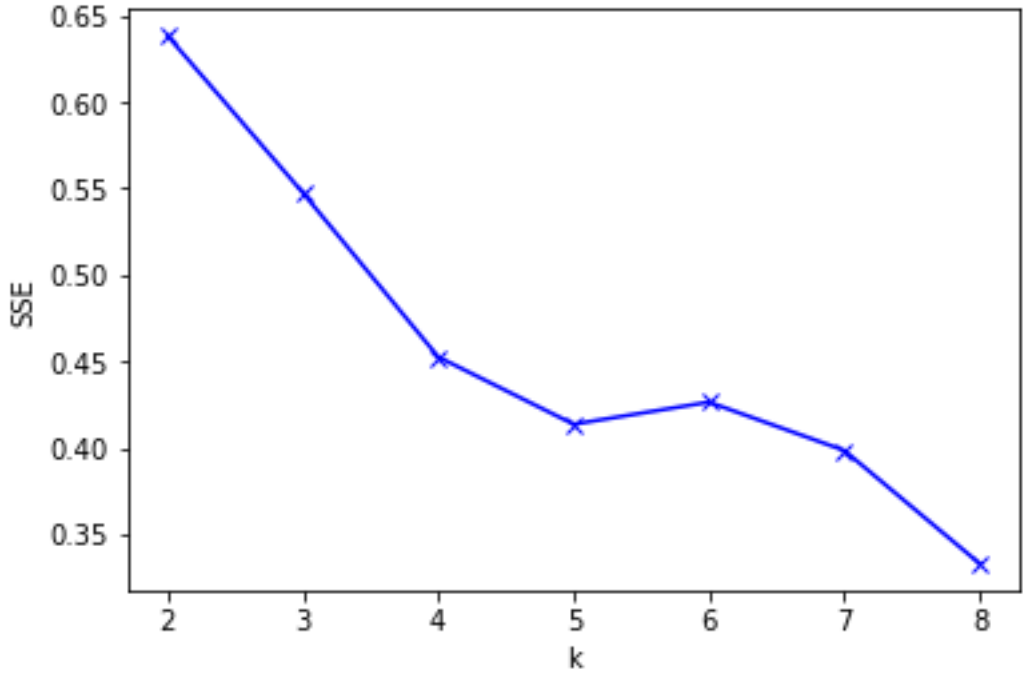

图 1 k 与 SSE 的关系图

所以我们将高钾玻璃分成两个亚类（具体划分结果见高钾一类. xlsx 和高钾二类.xlsx），用聚类算法 K-means 得出其质心为：

二氧化硅(SiO_2)	氧化钠(Na_2O)	氧化钾(K_2O)	氧化钙(CaO)	氧化镁(MgO)	氧化铝(Al_2O_3)	氧化铁(Fe_2O_3)	氧化铜(CuO)	氧化铅(PbO)	氧化钡(BaO)	五氧化二磷(P_2O_5)	氧化锶(SrO)	氧化锡(SnO_2)	二氧化硫(SO_2)
63.2	1.242727	10.04636	5.206364	1.097273	6.299091	2.080909	2.522727	0.37	0.473636	1.286364	0.039091	0	0.110909
89.66333	0	1.985556	1.326667	0.436667	2.764444	0.44	1.492222	0.138889	0.218889	0.533333	0.007778	0.262222	0

可以看出二氧化硅含量低、氧化钾含量高、氧化钙含量高、氧化铝含量高、氧化铁含量高的为高钾一类；二氧化硅含量高、氧化钾含量低、氧化钙含量低、氧化铝含量低、氧化铁含量低的为高钾二类，如图 2 所示

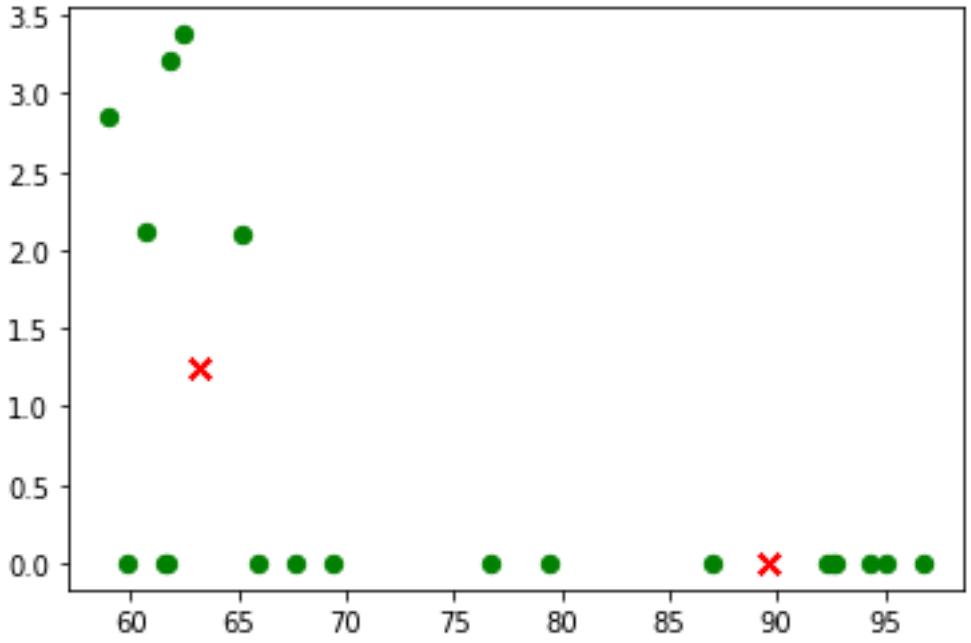

图 2 高钾玻璃聚类效果图

对于铅钡玻璃，我们将附件表单 2 里所有属于高钾玻璃的化学成分的数据导入到铅钡.xlsx 表里，然后采用手肘法得出 k 值，如图 3 所示，可以观察出 k 值为 2。

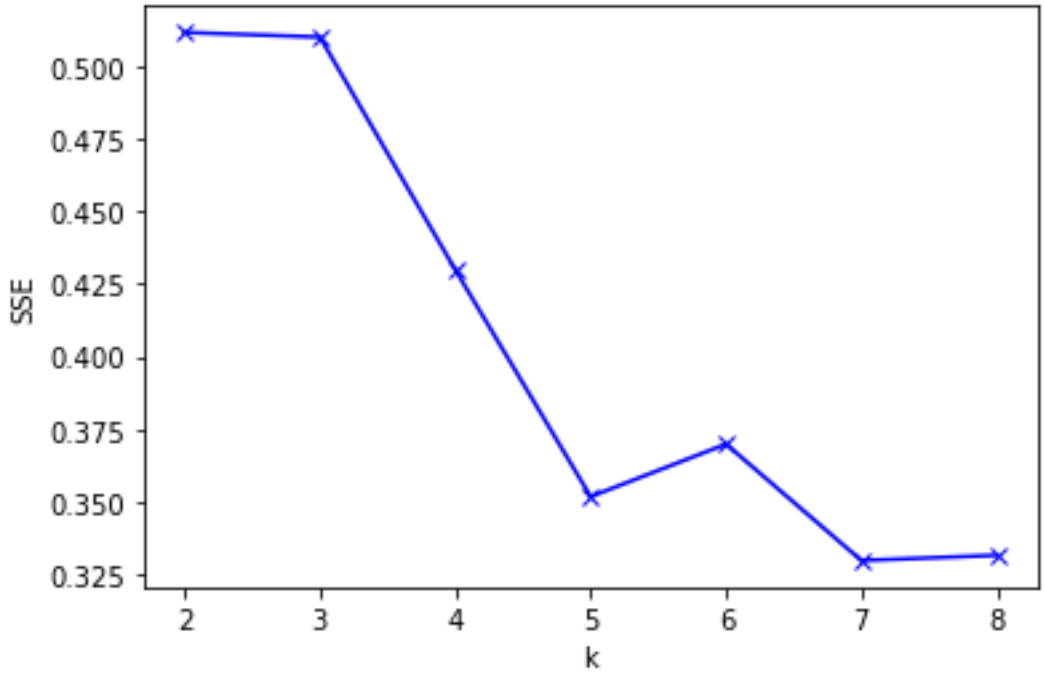

图 3 k 与 SSE 的关系图

所以我们将铅钡玻璃分成两个亚类（具体划分结果见铅钡一类. xlsx 和铅钡二类.xlsx），用聚类算法 K-means 得出其质心为：

二氧化硅(SiO_2)	氧化钠(Na_2O)	氧化钾(K_2O)	氧化钙(CaO)	氧化镁(MgO)	氧化铝(Al_2O_3)	氧化铁(Fe_2O_3)	氧化铜(CuO)	氧化铅(PbO)	氧化钡(BaO)	五氧化二磷(P_2O_5)	氧化锶(SrO)	氧化锡(SnO_2)	二氧化硫(SO_2)
24. 91464	0. 172143	0. 162857	2. 730714	0. 601786	2. 620357	0. 679643	2. 415357	43. 44821	12. 37679	4. 916429	0. 441786	0. 046786	1. 268571
57. 49	1. 880952	0. 187619	1. 142381	0. 70381	5. 06381	0. 624286	1. 165238	19. 88381	7. 975238	1. 127619	0. 222857	0. 073333	0. 174286

可以看出二氧化硅含量高、氧化铅高、氧化钡高的为铅钡一类；二氧化硅含量

低、氧化铅低、氧化钡低的为铅钡二类，如图 4 所示

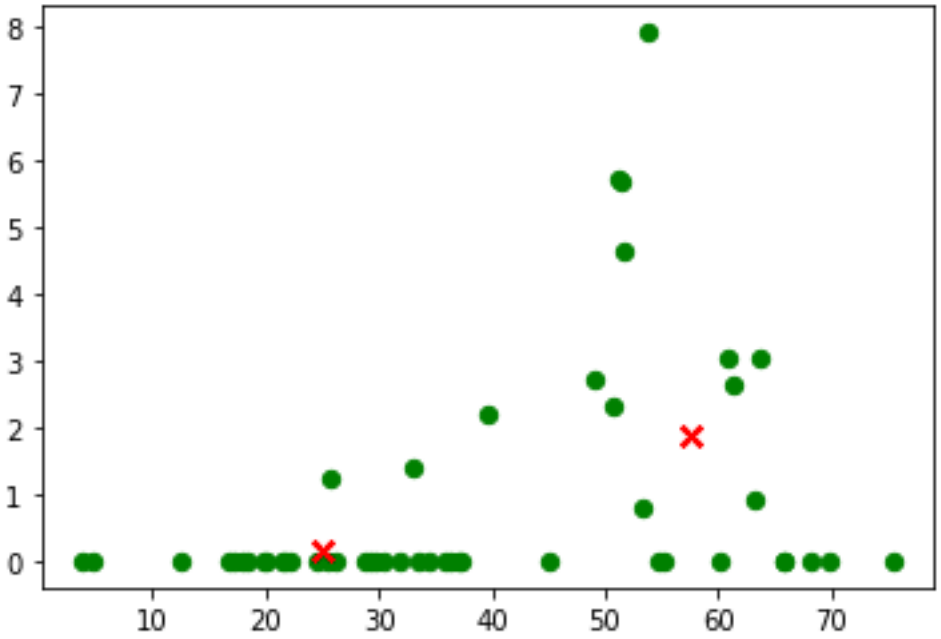

图 4 铅钡玻璃聚类效果图

5. 3 问题三模型的建立与求解

我们根据问题 2 所得出的分类规律来判断表单 3 中文物的类型，然后将其化学成分数据导入到高钾.xlsx 或者铅钡.xlsx 表中，用 K-means 聚类算法得出文物的所属亚类。

（1）对于文物 A1，我们可以观察出其二氧化硅含量高，氧化锶含量低，氧化铅含量低，氧化钡含量低，且文物表面无风化，所以文物 A1 属于高钾玻璃，将其化学成分导入高钾.xlsx 表里，用 K-means 聚类算法得出其属于高钾二类。

（2）对于文物 A2，我们可以观察出其二氧化硅含量低，氧化铅含量高，且文物表面风化，所以文物 A2 属于铅钡玻璃，将其化学成分导入铅钡.xlsx 表里，用 K-means 聚类算法得出其属于铅钡一类。

（3）对于文物 A3，我们可以观察出其二氧化硅含量低，氧化铅含量高，氧化锶含量高，且文物表面无风化，所以文物 A3 属于铅钡玻璃，将其化学成分导入铅钡.xlsx 表里，用 K-means 聚类算法得出其属于铅钡一类。

（4）对于文物 A4，我们可以观察出其二氧化硅含量低，氧化铅含量高，氧化锶含量高，且文物表面无风化，所以文物 A4 属于铅钡玻璃，将其化学成分导入铅钡.xlsx 表里，用 K-means 聚类算法得出其属于铅钡一类。

（5）对于文物 A5，我们可以观察出其二氧化硅虽然含量不低，但是其氧化铅含量高，氧化锶含量高，而且文物表面风化，所以我们得出文物 A5 属于铅钡玻璃，将其化学成分导入铅钡.xlsx 表里，用 K-means 聚类算法得出其属于铅钡二类。

（6）对于文物 A6，我们可以观察出其二氧化硅含量很高，氧化锶含量低，氧化铅含量低，氧化钡含量低，且文物表面风化，所以文物 A6 属于高钾玻璃，将其化学成分导入高钾.xlsx 表里，用 K-means 聚类算法得出其属于高钾二类。

（7）对于文物 A7，我们可以观察出其二氧化硅含量很高，氧化锶含量低，氧化铅含量低，氧化钡含量低，且文物表面风化，所以文物 A7 属于高钾玻璃，将其化学成分导入高钾.xlsx 表里，用 K-means 聚类算法得出其属于高钾二类。

（8）对于文物 A8，我们可以观察出其二氧化硅虽然含量不低，但是其氧化铅含量高，氧化锶含量高，而且文物表面无风化，所以我们得出文物 A8 属于铅钡玻璃，将其化学成分导入铅钡.xlsx 表里，用 K-means 聚类算法得出其属于铅钡二类。

5.4 问题四模型的建立与求解

5.4.1 化学成分关联关系

（1）灰色关联分析模型概述

灰色关联分析是一种多因素统计方法，其基本思想是通过计算主因子序列和每个行为因子序列之间的灰色关联度，来判断因子之间关系的强度、大小和顺序。主因子序列和行为因子序列之间的灰色关联度越大，则它们的关系越紧密，行为因子序列对主因子序列的影响越大，反之亦然。

（2）灰色关联分析模型的基本步骤

① 确定反映系统行为特征的参考序列$X^{(0)}$和影响系统行为的比较序列 $X^{(m)}$：

其中反映系统行为特征的数据序列为参考序列为：

$$X_0 = \{X_0(1), X_0(2), \dots, X_0(n)\} \tag{7}$$

影响系统行为的因素组成的数据序列为比较序列为：

$$\begin{gathered} X1 = X1(1), X1(2), \dots, X1(n) \\ X2 = X2(1), X2(2), \dots, X2(n) \\ \vdots \\ Xm = Xm(1), Xm(2), \dots, Xm(n) \end{gathered} \tag{8}$$

② 求各序列的初值像(进行无量纲化处理)。令

$$X' = X_i/X_i(1) = \left\{X_i'(1), X_i'(2), \dots, X_i'(n)\right\} \tag{9}$$

其中 i=0,1,2,...,m

得到 X'_0,X'_1,...,X'_m

③ 求 X_0与 X_i的初值像对应分量之差的绝对值序列。

记

$$\Delta_i(k) = |X_0'(k) - X_i'(k)|$$

$$\Delta_i = \left(\Delta_i(1), \Delta_i(2), \dots, \Delta_i(n)\right) \quad = 1,2, \dots, m; \mathrm{k} = 1,2, \dots, n\,\Delta$$

④ 求$\Delta_i(k) = |X_0'(k) - X_i'(k)|$的最小值与最大值。分别记为

$$\Delta_{min} = \min_i \min_k \quad \Delta_i(k)$$

$$\Delta_{max} = \max_i \max_k \quad \Delta_i(k)$$

⑤ 求关联系数$\xi_i(k)$

$$\xi_{0i}(k) = \frac{\Delta_{\min} + p\,\Delta_{\max}}{\Delta_i(k) + p\,\Delta_{\max}} \tag{10}$$

其中， p 为分辨系数，0< p <1，一般取 p=0.5。

⑥ 求关联度

$$\gamma_{0i} = \frac{1}{n}\sum_{k=1}^{n} \xi_{0i}(k) \tag{11}$$

⑦ 按 γ_{0i}大小排序 ，区分其关联程度的大小，若 γ_i值越大，说明其关联的程度越大；反之 γ_i值越小，则其关联程度越小

（3）对高钾玻璃与铅钡玻璃进行灰色关联分析

我们从问题一里知道对高钾玻璃影响较大的化学成分有二氧化硅、氧化钾、氧化钙、氧化铝、氧化铅、氧化钡。我们选择二氧化硅作为参考量，分析二氧化硅与其它四种变量之间的关联关系。如下是我们得出的灰色关联系数表、灰色关联系数图与灰色关联度。

表 17 灰色关联系数

氧化钾(K2O)	氧化钙(CaO)	氧化铝(Al2O3)	氧化铅(PbO)	氧化钡(BaO)
0.7602809344368593	0.8125613035 73525	0.8331868204294656	0.7602809344368593	0.7602809344368593
0.7699083045988695	0.7843283623 73761	0.8017419086215125	0.7552799631536 84	0.7552799631536 84
0.7744959129405158	0.7611622872511529	0.7792373166 29262	0.7516596872673141	0.7516596872673141
0.7822963539356007	0.7908845462544712	0.8087433245658563	0.7568004638019001	0.7568004638019001
0.7795867441283031	0.8453137991983967	0.9002220406 95988	0.7608711595041358	0.7608711595041358
0.7600914136580675	0.8056160429444978	0.8548970793195574	0.7600914136580675	0.7600914136580675
0.8687498416277615	0.8163406943399892	0.9980456680667565	0.8127445843952152	0.8127445843952152
0.9313769980599759	0.8802833032059701	0.9443114191 23889	0.9636326993334556	0.7722186989861995
0.7810519024409751	0.8215880330020441	0.9323190761215321	0.4537881533 2470157	0.3409143695 7374577
0.8695600045356412	0.7687441682475118	0.8992493763235201	0.8211345835975988	0.8211345835975988
0.8195025204549323	0.7492417144288764	0.8408147042386843	0.8318373391556242	0.8318373391556242
0.9564894529511926	0.8168086253678841	0.7249692644503801	0.9818429067293581	0.5607497373832873
0.9197294264023934	0.8419656548310177	0.7440490703633216	0.8929056229582687	0.6669684005860483
0.7710186879 42975	0.6892653778155543	0.8922222636909064	0.8383683560680383	0.8383683560680383
0.7851032517916294	0.7137558421 69888	0.7772182062504783	0.4137402543 8629225	0.8295992819578889

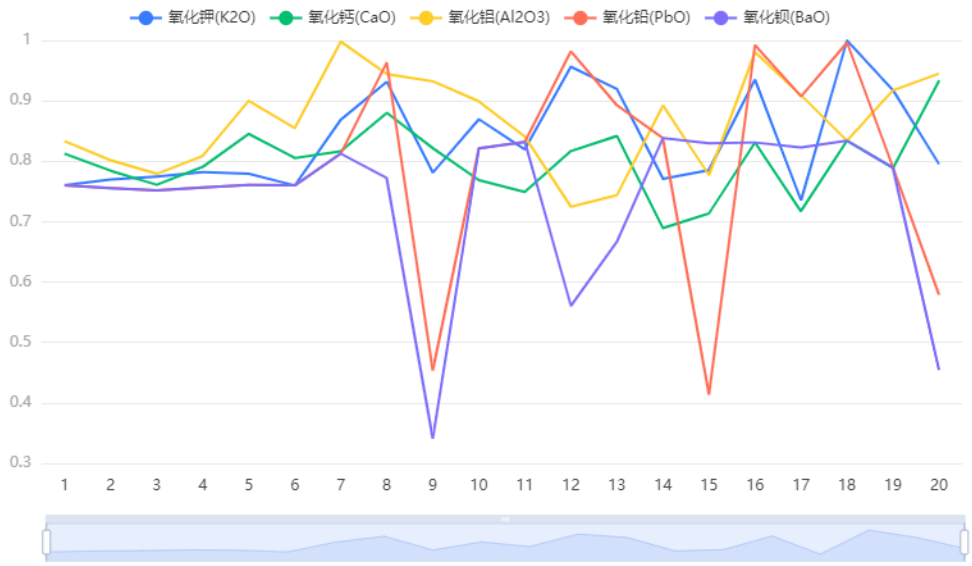

图 5 关联系数图

表 18 灰色关联度

评价项	关联度	排名
氧化铝(Al2O3)	0.866	1
氧化钾(K2O)	0.836	2
氧化钙(CaO)	0.8	3
氧化铅(PbO)	0.791	4
氧化钡(BaO)	0.738	5

由表 18 可知，氧化铝(Al2O3)与二氧化硅(SiO2)的关联度为 0.866，氧化钾(K2O)与二氧化硅(SiO2)的关联度为 0.836，氧化钙(CaO)与二氧化硅(SiO2)的关联度为 0.8，氧化铅(PbO)与二氧化硅(SiO2)的关联度为 0.791，氧化钡(BaO)与二氧化硅(SiO2)的关联度为 0.738，其中与二氧化硅(SiO2)关联度最大的是氧化铝(Al2O3)，与二氧化硅(SiO2)关联度最小的是氧化钡(BaO)。

我们从问题一里知道对铅钡玻璃影响较大的化学成分有二氧化硅、氧化钾、氧化钙、氧化铝、氧化铅、氧化钡。我们选择二氧化硅作为参考量，分析二氧化硅与其它四种变量之间的关联关系。如下是我们得出的灰色关联系数表、灰色关联系数图与灰色关联度。

表 19 灰色关联系数

氧化钾(K2O)	氧化钙(CaO)	氧化铝(Al2O3)	氧化铅(PbO)	氧化钡(BaO)
0.414799948710 09906	0.946632676535 4619	0.852920059975 5566	0.881989415112 2194	0.795982378271 565
0.875795901427 4178	0.947706585244 4708	0.960557170963 7205	0.914707397698 3789	0.596285514365 4625
0.969316328809 9425	0.716654393220 7181	0.952640321266 248	0.809989561754 6634	0.564596270755 6358
0.913642427630 9517	0.811130757114 0224	0.966217799410 197	0.973448834497 8582	0.873560210023 9714
0.826987769025 1056	0.845419585341 4415	0.945956106080 687	0.875057946168 9233	0.935830766367 4891
0.724365716040 9862	0.761465102440 4901	0.785085423880 2123	0.806380695279 4025	0.935674578381 4912
0.736398454170 686	0.785414847843 9731	0.822979318751 5071	0.913929205521 2151	0.845202657361 9863
0.877704373482 1945	0.950316644466 2708	0.920215475417 6421	0.906817879889 089	0.586047946186 308
0.621715736785 1769	0.725967397492 3551	0.942248812519 447	0.819688036422 5961	0.525083404549 8998
0.935839588348 5027	0.768268645302 9961	0.886141926404 7474	0.746183169626 1628	0.726911177615 5126
0.973857920697 6444	0.955012398796 4815	0.614186744067 3555	0.742839509110 2836	0.717044032034 7292
0.875216986311 1562	0.871204336493 3777	0.884210188630 6061	0.884900576162 8724	0.992043666391 0407
0.945996142191 1434	0.813091071089 5787	0.862558711002 6883	0.941321377639 7868	0.997054617580 6805
0.811328966189 3376	0.876369157436 3743	0.962181922442 0095	0.852435060569 1431	0.977788065980 8803
0.843767625133 4238	0.965397684903 3801	0.871410432320 7823	0.759065821762 3587	0.997463734482 8706

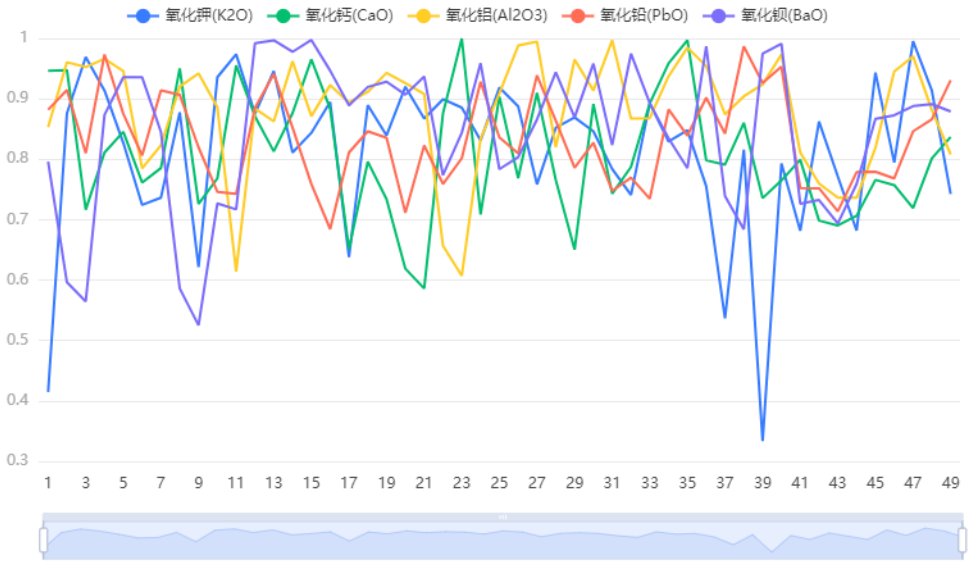

图 6 灰色关联系数图

表 20 灰色关联度

评价项	关联度	排名
氧化铝(Al2O3)	0.881	1
氧化钡(BaO)	0.845	2
氧化铅(PbO)	0.834	3
氧化钾(K2O)	0.814	4

由表 20 可知，氧化铝(Al2O3)与二氧化硅(SiO2)的关联度为 0.881，氧化钡(BaO)与二氧化硅(SiO2)的关联度为 0.845，氧化铅(PbO)与二氧化硅(SiO2)的关联度为 0.834，氧化钾(K2O)与二氧化硅(SiO2)的关联度为 0.814，氧化钙(CaO)与二氧化硅(SiO2)的关联度为 0.808，其中与二氧化硅(SiO2)关联度最大的是氧化铝(Al2O3)，与二氧化硅(SiO2)关联度最小的是氧化钙(CaO)。

5. 4. 2 化学成分关联关系的差异性

研究高钾玻璃和铅钡玻璃中二氧化硅（SiO2）与氧化铝（Al2O3）关联关系的差异性，使用配对样本 T 检验进行差异性分析。配对样本 T 检验是对用于检验配对设计实验中成对定量数据是否存在差异性的统计方法。

（1）正态性检验

对数据进行 Shapiro-Wilk 检验，查看其显著性：若呈现出显著性(P<0.05)，说明不符合正态分布，通常现实研究情况下很难满足检验，若其样本峰度绝对值小于 10 并且偏度绝对值小于 3，结合正态分布图可以描述为基本符合正态分布，亦或改用非参数检验；若不呈现出显著性(P<0.05)，说明符合正态分布。

Shapiro-Wilk 检验的理论

该检验的零检验是样本 x1,⋯,xn 来自于一个正态分布的母体。

这个检验的统计量是：

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{12}$$

其中

x(i) 用括号包含下标索引 i 的；不与 x 混淆,它是第 i 阶统计量，即样本中的第 i 个最小数

$\bar{x} = (x_1 + \cdots + x_n)/n$是样本的平均值。

常量 ai 通过公式[1]

$$(a_1, \ldots, a_n) = \frac{m^{\top} V^{-1}}{(m^{\top} V^{-1} V^{-1} m)^{1/2}} \tag{13}$$

$$m = (m_1, \ldots, m_n)^{\top}$$

其中 $m_1,\cdots,m_n$

是从一个标准的正态分布随机变量上采样的有序独立同分布的统计量的期望值。

V 是这些有序统计量的协方差。

表 21 配对差值正态性检验结果

变量名	样本量	平均值	标准差	偏度	峰度	S-W 检验
高钾玻璃二氧化硅和氧化铝	20	0.866	0.078	-0.135	-0.888	-------
铅钡玻璃二氧化硅和氧化铝	20	0.891	0.082	-2.266	6.535	-------
高钾玻璃二氧化硅和氧化铝配对铅钡玻璃二氧化硅和氧化铝	20	-0.025	0.108	0.661	0.371	0.942(0.259)

高钾玻璃二氧化硅和氧化铝配对铅钡玻璃二氧化硅和氧化铝样本 N＜5000，显著性 P 值为 0.259，水平上不呈现显著性，不能拒绝原假设，因此数据满足正态分布，其峰度（0.371）绝对值小于 10 并且偏度（0.661）绝对值小于 3。

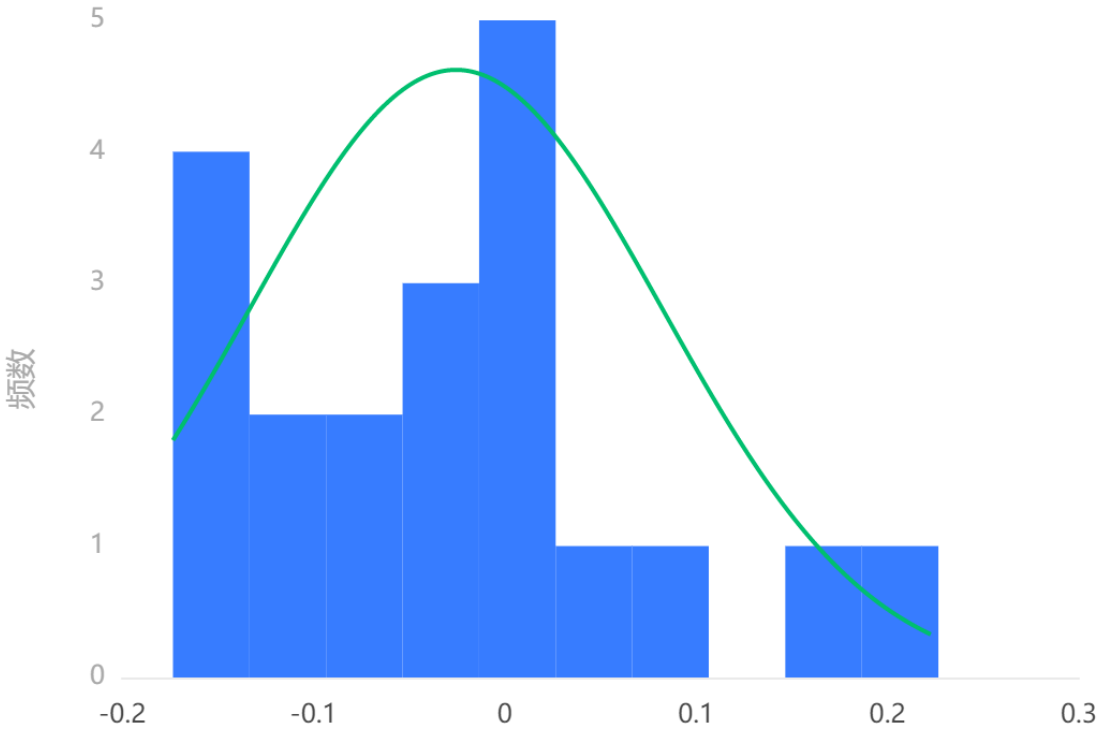

图 7 高钾玻璃二氧化硅和氧化铝-铅钡玻璃二氧化硅和氧化铝正态性检验直方图

（2）配对样本 T 检验

在通过正态性检验后，可以通过配对 T 检验判断 P 值是否呈现出显著性(P<0.05)，若呈显著性，根据均值与检验值进行差异分析，描述差异大小。

表 22 配对样本 T 检验结果

配对变量	平均值±标准差			t	df	P	Cohen's d
	配对 1	配对 2	配对差值（配对 1-配对 2）				
高钾玻璃二氧化硅和氧化铝配对铅钡玻璃二氧化硅和氧化铝	0.866±0.078	0.891±0.082	-0.025±-0.004	-1.056	19	0.304	0.236

配对样本 T 检验的结果显示，基于变量高钾玻璃二氧化硅和氧化铝配对铅钡玻璃二氧化硅和氧化铝，显著性 P 值为 0.304，水平上不呈现显著性，不能拒绝原假设，因此字高钾玻璃二氧化硅和氧化铝配对铅钡玻璃二氧化硅和氧化铝之间不存在显著性差异。其差异幅度 Cohen's d 值为：0.236，差异幅度较小。

六、 模型的评价、改进

6.1 模型的优点

1. 在问题一中，我们采用了斯皮尔曼相关性检验和卡方检验，分别从相关性和差异性两方面研究玻璃文物表面风化和类型、纹饰、颜色的关系。

2. 在问题二中，我们对二元 logistic 回归模型进行了一定的改进，分别研究了每一种化学成分在回归方程中的显著性，从而排除了众多的化学成分，提取出了真正影响玻璃文物分类的特征变量。

3. K-means 聚类模型快速、简单，对大数据集有较高的效率并且是可伸缩性的，其时间复杂度近于线性，而且适合挖掘大规模数据集。

6.2 模型的缺点

1. K-means 聚类模型过于简单从而导致收初始值和离群点的影响，每次结果都不稳定，聚类中心不一定属于数据集。

2. 二元 logistic 回归模型更适合市场类型的二元决策问题，对于本题的玻璃文物分类使用起来较为勉强。

七、 参考文献

[1] 贾晓芬,郭永存,黄友锐,等. 基于斯皮尔曼等级相关性的彩色图像椒盐噪点检测算法[J]. 中国科学技术大学学报,2019,49(1):63-70.

[2] 孙楠. 卡方检验在笔迹学中的应用[J]. 广东公安科技,2020,28(2):27-29.

[3] 彭佳琪,徐璐. 武汉市在校生垃圾分类认知和行为调查研究 ——基于二元 Logistic 回归模型[J]. 科教导刊,2022(7):151-154.

[4] 杨俊闯,赵超. K-Means 聚类算法研究综述[J]. 计算机工程与应用,2019,55(23):7-14,63.

附录

附录 1

介绍：支撑材料文件列表

结论数据文件夹：第一题的预测结果，第二题的二元逻辑回归结果，第二题的亚类划分结果：

1. Result1.xlsx 为第一题预测结果
2. Result2.xlsx 第二题二元逻辑回归结果
3. 高钾一类.xlsx、高钾二类.xlsx、铅钡一类.xlsx、铅钡二类.xlsx 为四个亚类的划分结果

代码文件夹：所有 python 的程序源代码及所需的数据文件，具体如下：

1. Num1_relevance.py 和附件 1.xlsx 是第一题斯皮尔曼关联性分析的代码及数据
2. Num1_difference.py 和附件 1.xlsx 是第一题卡方检验差异性分析的代码及数据
3. Num2.py、高钾.xlsx 和铅钡.xlsx 是第二题聚类分析寻找亚类的代码及数据

附录 2

介绍：问题一分析相关性与差异性（python 代码） 文件名称：Num1_relevance.py

```
import os
import numpy as np
import pandas as pd
from pathlib import Path
from  scipy.stats import chi2_contingency
filename = Path("D:\数模\C 题\附件 1.xlsx")
# read the file 使用哪列：usecols 参数从 0 开始
df= pd.read_excel(filename,sheet_name='表单 1')
a=pd.crosstab(df['表面风化'], df['纹饰'])
print(a)
kf_datad = np.array(a)
kf = chi2_contingency(kf_datad)
print('皮尔逊卡方=%.4f, 渐进显著性=%.4f, 自由度=%i expected_frep=%s'%kf)
b=pd.crosstab(df['表面风化'], df['颜色'])
print(b)
kf_datad = np.array(b)
kf = chi2_contingency(kf_datad)
print('皮尔逊卡方=%.4f, 渐进显著性=%.4f, 自由度=%i expected_frep=%s'%kf)
c=pd.crosstab(df['表面风化'], df['类型'])
print(c)
kf_datad = np.array(c)
kf = chi2_contingency(kf_datad)
print('皮尔逊卡方=%.4f, 渐进显著性=%.4f, 自由度=%i expected_frep=%s'%kf)
```

附录 3

介绍：问题一分析相关性与差异性（python 代码）文件名称：Num1_difference.py

```
import pandas as pd
```

```
    import numpy as np
    import scipy

    def excel_one_line_to_list():
        X1 = pd.read_excel("D:\ 数 模 \C 题 \ 附 件 1.xlsx",sheet_name=' 表 单 1',usecols=[4],
                           names=None) # 读取项目名称列,不要列名
        X1 = np.array(X1)
        X1 = X1.tolist()
        X1 = pd.Series(X1)
        X1 = X1.astype('str')
        Y1 = pd.read_excel("D:\ 数 模 \C 题 \ 附 件 1.xlsx",sheet_name=' 表 单 1',usecols=[2],
                           names=None) # 读取项目名称列,不要列名
        Y1 = np.array(Y1)
        Y1 = Y1.tolist()
        Y1 = pd.Series(Y1)
        Y1 = Y1.astype('str')
        # 处理数据删除 Nan
        x1 = X1.dropna()
        y1 = Y1.dropna()
        n = x1.count()
        x1.index = np.arange(n)
        y1.index = np.arange(n)
        # 分部计算
        d = (x1.sort_values().index - y1.sort_values().index) ** 2
        dd = d.to_series().sum()
        p = 1 - n * dd / (n * (n ** 2 - 1))
        # s.corr()函数计算
        r = x1.corr(y1, method='spearman')
        print(round(r,3), p)  # 0.942857142857143 0.9428571428571428
        print(scipy.stats.spearmanr(X1, Y1))
    if __name__ == '__main__':
  excel_one_line_to_list()
```

附录 4

介绍：问题二聚类分析（python 代码） 文件名称：Num2.py

```
import random
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# 计算欧拉距离
def calcDis(dataSet, centroids, k):
  clalist=[]
  for data in dataSet:
    diff = np.tile(data, (k, 1)) - centroids  #相减   (np.tile(a,(2,1))就是把 a 先沿 x 轴复
```

```
制 1 倍，即没有复制，仍然是 [0,1,2]。 再把结果沿 y 方向复制 2 倍得到 array([[0,1,2],[0,1,2]]))
        squaredDiff = diff ** 2    #平方
        squaredDist = np.sum(squaredDiff, axis=1)  #和 (axis=1 表示行)
        distance = squaredDist ** 0.5  #开根号
        clalist.append(distance)
    clalist = np.array(clalist)  #返回一个每个点到质点的距离 len(dateSet)*k 的数组
    return clalist

# 计算质心
def classify(dataSet, centroids, k):
    # 计算样本到质心的距离
    clalist = calcDis(dataSet, centroids, k)
    # 分组并计算新的质心
    minDistIndices = np.argmin(clalist, axis=1)    #axis=1 表示求出每行的最小值的下标
    newCentroids = pd.DataFrame(dataSet).groupby(minDistIndices).mean()
#DataFramte(dataSet)对 DataSet 分组，groupby(min)按照 min 进行统计分类，mean()对分类结果求均值
    newCentroids = newCentroids.values

    # 计算变化量
    changed = newCentroids - centroids

    return changed, newCentroids

# 使用 k-means 分类
def kmeans(dataSet, k):
    # 随机取质心
    centroids = random.sample(dataSet, k)

    # 更新质心 直到变化量全为 0
    changed, newCentroids = classify(dataSet, centroids, k)
    while np.any(changed != 0):
        changed, newCentroids = classify(dataSet, newCentroids, k)

    centroids = sorted(newCentroids.tolist())   #tolist()将矩阵转换成列表 sorted()排序

    # 根据质心计算每个集群
    cluster = []
    clalist = calcDis(dataSet, centroids, k) #调用欧拉距离
    minDistIndices = np.argmin(clalist, axis=1)
    for i in range(k):
        cluster.append([])
    for i, j in enumerate(minDistIndices):   #enymerate()可同时遍历索引和遍历元素
        cluster[j].append(dataSet[i])

    return centroids, cluster
from sklearn.metrics import silhouette_score
```

```
from sklearn.cluster import KMeans
#手肘法算出 k 值
def get_silhouette_K(data, range_K):
    K = range(2, range_K)
    Scores = []
    for k in K:
        kmeans = KMeans(n_clusters=k)
        kmeans.fit(data)
        Scores.append(silhouette_score(data, kmeans.labels_, metric='euclidean'))
    max_idx = Scores.index(max(Scores))
    best_k = K[max_idx]
    plt.plot(K, Scores, 'bx-')
    plt.xlabel('k')
    plt.ylabel('SSE')
    plt.show()
    return best_k
# 创建数据集
def createDataSet():
    df= pd.read_excel("D:\数模\C 题\高钾.xlsx",header=None)
    #df= pd.read_excel("D:\数模\C 题\铅钡.xlsx",header=None)
    t=np.array(df)
    Temp_List = t.tolist()
    return Temp_List

if __name__=='__main__':
    dataset = createDataSet()
    k=get_silhouette_K(dataset,9)
    centroids, cluster = kmeans(dataset,k)
    print('质心为：%s' % centroids)
    print('集群为：%s' % cluster)
    #结果导出到 excel
    '''
    df = pd.DataFrame(centroids)
    df.to_excel("D:\数模\C 题\高钾质心.xlsx",sheet_name='高钾质心', index=False)
    df = pd.DataFrame(cluster[0])
    df.to_excel("D:\数模\C 题\高钾一类.xlsx",sheet_name='高钾一类', index=False)
    df = pd.DataFrame(cluster[1])
    df.to_excel("D:\数模\C 题\高钾二类.xlsx",sheet_name='高钾二类', index=False)
    '''
    '''
    df = pd.DataFrame(centroids)
    df.to_excel("D:\数模\C 题\铅钡质心.xlsx",sheet_name='铅钡质心', index=False)
    df = pd.DataFrame(cluster[0])
    df.to_excel("D:\数模\C 题\铅钡一类.xlsx",sheet_name='铅钡一类', index=False)
    df = pd.DataFrame(cluster[1])
    df.to_excel("D:\数模\C 题\铅钡二类.xlsx",sheet_name='铅钡二类', index=False)
    '''
```

```
    for i in cluster:
        for j in range(len(i)):
            plt.scatter(i[j][0],i[j][1], marker = 'o',color = 'green', s = 40 ,label = '原始点')
                                            # 记号形状       颜色      点的大小      设置标签
            for k in range(len(centroids)):
                plt.scatter(centroids[k][0],centroids[k][1],marker='x',color='red',s=50,label=' 质
心')
    plt.show()
```